

Probability Theory

Notes taken by Yeheli Fomberg <yp@campus.technion.ac.il>

Based on lectures by and lecture notes by Ron Rosenthal

Winter 2025

These notes are not endorsed by the lecturers. I have revised them outside lectures to incorporate supplementary explanations, clarifications, and material for fun. While I have strived for accuracy, any errors or misinterpretations are most likely mine.

Contents

1	Probability spaces	4
1.1	Introduction	4
1.2	σ -algebras	5
1.3	Properties of probability functions	7
2	Conditional probability	9
2.1	Conditional probability	9
2.2	The law of total probability	9
2.3	Bayes' theorem	11
3	Independence and repeating experiments	13
3.1	Independence of events	13
3.2	Bernoulli Trials	14
3.3	Random walks	15
3.4	Infinite coin flips	16
4	Random variables	17
4.1	Random variables	17
4.2	Coupon collector	18
4.3	Cumulative distribution function	19
4.4	Binomial distribution	21
4.5	Geometric distribution	22
4.6	Divergence of the sum of the reciprocals of the primes	23
4.7	Continuous and absolutely continuous random variables	25
4.8	Uniform distribution	26
4.9	Normal distribution	27
4.10	Mixed random variables	27
4.11	Measurable spaces and measurable functions	28
4.12	Simulation of random variables	30
5	Expectation, variance and moments	31
5.1	Expectation	31
5.2	Variance and standard deviation	33
5.3	Moments	35
6	Important inequalities	36
6.1	Jensen's inequality	36
6.2	Markov's inequality	37
6.3	Chebyshev's inequality	38
6.4	The weak law of large numbers	39
6.5	The Weierstrass approximation theorem	40
7	Bertrand's paradox	41
7.1	First solution	41
7.2	Second solution	41
7.3	Third solution	42
7.4	Conclusion	42
8	Multivariate random variables	43
8.1	Multivariate random variables	43
8.2	Theory of random vectors	45

9	Expectation of random vectors, covariance and correlation	52
9.1	Expectation	52
9.2	Covariance and correlation	53
9.3	Indicators and expectation	56
9.4	Cauchy distribution	57
10	Conditional probability distribution and expectation	58
10.1	The discrete case	58
10.2	The absolutely continuous case	60
11	The Borel–Cantelli lemmas	63
11.1	Set theoretic preparation	63
11.2	Fatou’s lemma	63
11.3	The Borel–Cantelli lemmas	65
12	The law of large numbers	67
12.1	Convergence of random variables	67
12.2	The law of large numbers	69
12.3	The law of truly large numbers	71
13	The central limit theorem	73
13.1	Introduction to the theorem	73
13.2	Preliminaries for the proof	75
13.3	Miscellaneous remarks and definitions	76
14	Random walks	78
14.1	Simple symmetric random walks	78
14.2	One-dimensional SSRW	79
14.3	Non-symmetric simple random walk	80

1 Probability spaces

1.1 Introduction

Before diving in into the definition of a probability space, the main object of this course, we must note that this course is an introductory course in probability theory, which means we don't have the tools from measure theory to formalize probability. Thus, some proofs will be omitted, and we will also need to formalize discrete and continuous probability theory separately.

Paradox 1.1 (Bertrand's paradox). Consider an equilateral triangle inscribed inside a circle. Suppose a chord of the circle is chosen at random. What is the probability that the chord is longer than a side of the triangle?

Remark 1.1. We can ponder about this paradox for a while, but Bertrand himself came up with three solutions, each with a different answer. The main difference in his methods lies in the way in which we choose the chords. More on that in Section 7.

Definition 1.1 (Sample space). The sample space of an experiment, is a set Ω which contains all the possible outcomes of the experiment.

Remark 1.2. A good thing to note, is that we can choose different sample spaces for the same experiment. For example, if the experiment consists of rolling two dice, and we want to check for the sum of the results, we can set either $\Omega = \{1, 2, 3, 4, 5, 6\}^2$, for the result of each dice, or $\Omega = \{1, 2, \dots, 11, 12\}$ for the sum of the results of the dice.

Definition 1.2 (Probability space, intuitive definition). A discrete probability space is a pair $(\Omega, \mathbf{P})$, where Ω is a countable sample set, and $\mathbf{P}: \Omega \rightarrow [0, 1]$ is a function such that $\sum_{\omega \in \Omega} \mathbf{P}(\omega) = 1$. Intuitively, we say that $\mathbf{P}(\omega)$ represents the probability that ω will happen.

Definition 1.3 (Event). A subset of the sample space $A \subseteq \Omega$ is called an event. We define the probability of an event as follows:

$$\mathbf{P}(A) := \sum_{\omega \in A} \mathbf{P}(\omega).$$

Remark 1.3. Here are a few properties of events and probability functions we can immediately verify:

- (1) $\mathbf{P}(\Omega) = 1$
- (2) $\mathbf{P}(\emptyset) = 0$
- (3) For $A \subset \Omega$ we have $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$
- (4) If $\{A_n\}_{n=1}^N$ are disjoint sets then

$$\mathbf{P}\left(\bigcup_{n=1}^N A_n\right) = \sum_{n=1}^N \mathbf{P}(A_n).$$

- (5) If $\{A_n\}_{n=1}^\infty$ is a sequence of pairwise disjoint sets then

$$\mathbf{P}\left(\bigcup_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty \mathbf{P}(A_n).$$

Definition 1.4 (Uniform probability function). In a finite probability space $(\Omega, \mathbf{P})$ we say that the probability function is uniform if for every $\omega \in \Omega$ we have $\mathbf{P}(\omega) = \frac{1}{|\Omega|}$.

Example 1.1 (The intuitive definition fails). We now proceed to consider an experiment in which we choose a direction in $\mathbb{R}^2$ at random and write it on a blank page. We define the sample space to be

$$\Omega := S^1 = \left\{ e^{i\theta} \mid \theta \in [0, 2\pi) \right\}.$$

A natural question to ask, is if we can define a uniform probability function in the sense that for any arc $[a, b] \subset S^1$ we have $\mathbf{P}([a, b]) = \frac{b-a}{2\pi}$. The answer is that with the definition we have worked with so far, we can't. Assume, by way of contradiction, that there was such a probability function $\mathbf{P}$. Then $\mathbf{P}(\{a\}) = 0$ for all $a \in \Omega$, and thus we have that

$$\mathbf{P}(\Omega) = \sum_{\omega \in \Omega} \mathbf{P}(\omega) = 0 \neq 1,$$

which is a contradiction. In order to solve this problem, we may try to define a new function $\mathbf{P}: 2^\Omega \rightarrow [0, 1]$ that will directly assign each event its probability directly, but unfortunately for us, such a function, which also satisfies the desired properties of a probability function, does not exist. The full proof for this is given in the course “real functions”, and will not be discussed here. However, we can give a proof, under the assumption of the following lemma.

Lemma 1.1. *There exists a set $E \subset S^1$ such that for any rational number $q \in (0, 2\pi) \cap \mathbb{Q}$ we have $e^{iq}E \cap E = \emptyset$.*

Indeed we see that

$$1 = \mathbf{P}(\Omega) = \mathbf{P}\left(\bigcup_{q \in [0, 2\pi) \cap \mathbb{Q}} e^{iq}E\right) = \sum_{q \in [0, 2\pi) \cap \mathbb{Q}} \mathbf{P}(e^{iq}E) = \sum_{q \in [0, 2\pi) \cap \mathbb{Q}} \mathbf{P}(E)$$

and now we have a contradiction because if we set $\mathbf{P}(E) = \epsilon > 0$ then

$$1 = \sum_{q \in [0, 2\pi) \cap \mathbb{Q}} \epsilon,$$

and this is a contradiction for any $\epsilon > 0$. Of course we also get a contradiction if we set $\mathbf{P}(E) = 0$.

The classical solution to this problem, is to restrict the domain of the probability function to a certain subset of 2^Ω . We denote this new domain as $\mathcal{F} \subset 2^\Omega$. In order for $\mathbf{P}$ to keep its desired properties as a probability function, we must also require $\mathcal{F}$ to satisfy some properties.

1.2 σ -algebras

Definition 1.5 (σ -algebra). Let Ω be a set. We say that $\mathcal{F} \subset 2^\Omega$ is a σ -algebra of sets (sometimes called a σ -field, which is why we denote it by $\mathcal{F}$), if it satisfies the following properties:

- (1) $\Omega \in \mathcal{F}$.
- (2) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
- (3) If $\{A_n\}_{n=1}^\infty \subset \mathcal{F}$, then $\bigcup_{n=1}^\infty A_n \in \mathcal{F}$.

Definition 1.6 (Probability space). A probability space is a triplet $(\Omega, \mathcal{F}, \mathbf{P})$ such that Ω is a set, $\mathcal{F}$ is a σ -algebra of Ω , and $\mathbf{P}: \mathcal{F} \rightarrow [0, 1]$ is a probability function that satisfies:

- (1) $\mathbf{P}(\Omega) = 1$
- (2) If $\{A_n\}_{n=1}^\infty \subset \mathcal{F}$ are disjoint, then $\mathbf{P}\left(\bigcup_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty \mathbf{P}(A_n)$.

Definition 1.7 (Event). An event in the formal setting, is an element of $\mathcal{F}$.

Remark 1.4. If not stated otherwise, we always assume $\mathcal{F} = 2^\Omega$, when Ω is finite.

Proposition 1.2. *There exists a σ -algebra $\mathfrak{B}$ of $\Omega = S^1$, and a unique function $\mathbf{P}: \rightarrow [0, 1]$ such that $(\Omega, \mathfrak{B}, \mathbf{P})$ is a probability space and $\mathbf{P}$ is invariant to spinning on the sphere.*

Definition 1.8 (Algebra of sets). A set $\mathcal{C} \subset 2^\Omega$ is said to be an algebra of sets if it satisfies the following properties:

- (1) $\Omega \in \mathcal{C}$.
- (2) If $A \in \mathcal{C}$, then $A^c \in \mathcal{C}$.
- (3) if $A, B \in \mathcal{C}$, then $A \cup B \in \mathcal{C}$.

Remark 1.5. Any algebra $\mathcal{C}$ is closed under finite unions and finite intersections. It is also worth noting that we always have $\emptyset \in \mathcal{C}$, and that if $A, B \in \mathcal{C}$, then $A \setminus B \in \mathcal{C}$. We can also notice that any σ -algebra is closed under countable intersections, and that every σ -algebra is in particular also an algebra.

Example 1.2. If Ω is a set, and $A \subset \Omega$, then both 2^Ω and $\{\emptyset, A, A^c, \Omega\}$ are σ -algebras.

Example 1.3. Given a set Ω , the smallest σ -algebra of Ω is $\{\emptyset, \Omega\}$ which is called the trivial σ -algebra.

Proposition 1.3. *Let $(\mathcal{F}_\alpha)_{\alpha \in I}$ be a family of σ -algebras, then $\cap_{\alpha \in I} \mathcal{F}_\alpha$ is a σ -algebra.*

Proof. Obvious. □

Definition 1.9 (Minimal sigma algebra). Let Ω be a set, and let $H \subset 2^\Omega$ be a family of subsets. Then we define the minimal sigma algebra that contains H , denoted $\sigma(H)$, to be the intersection of all the σ -algebras that contain all elements in H . Notice that the intersection is never empty because 2^Ω is a σ -algebra that will always contain the elements of H . We say that $\sigma(H)$ is the σ -algebra generated by H .

Example 1.4 (Borel's σ -algebra). One of the most important minimal σ -algebras, is Borel's σ -algebra defined on $\mathbb{R}$. It is defined as such:

$$\mathfrak{B} = \mathfrak{B}(\mathbb{R}) := \sigma \left(\{ (a, b) \mid a < b \} \right).$$

That is, the smallest σ -algebra that contains all the open intervals in $\mathbb{R}$. Similarly, we can define it on the space $\mathbb{R}^d$ as follows:

$$\mathfrak{B}_d = \mathfrak{B}(\mathbb{R}^d) := \sigma \left(\left\{ \prod_{i=1}^d (a_i, b_i) \mid a_i < b_i \right\} \right).$$

In general, given a topological space (X, τ) , we define $\mathfrak{B}(X) = \sigma(\tau)$. It can be shown that this definition coincides with the definitions we just gave for $\mathfrak{B}$ and $\mathfrak{B}_d$.

Remark 1.6. It may seem that the notion of σ -algebras is very abstract and unnatural. But notice that its definition captures *exactly* what we need when we want to measure probability (or measure anything in general). When we require $\mathbf{P}(\Omega) = 1$ (captured by the first property) and say that we can measure the probability of A , we must also allow A^c to be measurable (captured by the second property). Moreover, when we have multiple disjoint events, for example when we roll a dice and we can measure the probability of it landing on 1, and the probability of it landing on 2, we also want to be able to measure the probability it landing on either 1 or 2 (captured by property 3). And we also want this to be true in a more general case when we have a countably infinite amount of disjoint events (this is why we use a σ -algebra and not just an algebra).

Theorem 1.4 (Carathéodory's theorem). *Let Ω be a set, let $\mathcal{G}$ be an algebra of sets of Ω . If $\hat{\mathbf{P}}: \mathcal{G} \rightarrow [0, 1]$ is a function that satisfies $\hat{\mathbf{P}}(\Omega) = 1$, and for each sequence of pairwise disjoint sets $\{A_n\}_{n=1}^{\infty}$ that*

$$\hat{\mathbf{P}}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \hat{\mathbf{P}}(A_n),$$

then there exists a single extension $\mathbf{P}: \sigma(\mathcal{G}) \rightarrow [0, 1]$ to $\hat{\mathbf{P}}: \mathcal{G} \rightarrow [0, 1]$, such that the triplet $(\Omega, \sigma(\mathcal{G}), \mathbf{P})$ is a probability space.

Proof. We will not give the proof of this in this course, but we will prove it in the course “real functions”. $\square$

Remark 1.7. Consider the problem from Example 1.1. In order to find a uniform probability function on it we first define the set $\mathcal{G}$ to be the set of all finite unions of intervals on S^1 . As it is closed under union of pairs, and complements, it is an algebra of sets. Next, define $\hat{\mathbf{P}}: \mathcal{G} \rightarrow [0, 1]$ as such:

$$\hat{\mathbf{P}}\left(\biguplus_{i=1}^N (a_i, b_i)\right) = \sum_{i=1}^N \frac{b_i - a_i}{2\pi},$$

We can see that $\hat{\mathbf{P}}$ satisfies the conditions in Theorem 1.4 and thus there exists an extension $\mathbf{P}$ defined on the σ -algebra $\mathcal{B} = \sigma(\mathcal{G})$ which is also called the Borel σ -algebra of S^1 . We have that $(\Omega, \mathcal{B}, \mathbf{P})$ is a probability space and $\mathbf{P}$ is uniform probability function on S^1 .

1.3 Properties of probability functions

Proposition 1.5. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Then*

(1) $\mathbf{P}(\emptyset) = 0$.

(2) *If $\{A_n\}_{n=1}^N \subset \mathcal{F}$ are disjoint sets then $\bigcup_{n=1}^N A_n \in \mathcal{F}$ and*

$$\mathbf{P}\left(\bigcup_{n=1}^N A_n\right) = \sum_{n=1}^N \mathbf{P}(A_n).$$

(3) *For every $A \in \mathcal{F}$ we have $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$.*

(4) *If $A, B \in \mathcal{F}$ and $A \subset B$, then $\mathbf{P}(B \setminus A) = \mathbf{P}(B) - \mathbf{P}(A)$ and thus $\mathbf{P}(A) \leq \mathbf{P}(B)$.*

(5) *If $A, B \in \mathcal{F}$, then*

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

Proposition 1.6. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Then*

(1) *If $(A_n)_{n=1}^{\infty} \subset \mathcal{F}$ is an increasing sequence of events, that is $A_1 \subset A_2 \subset A_3, \dots$, then*

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n).$$

(2) *If $(A_n)_{n=1}^{\infty} \subset \mathcal{F}$ is a decreasing sequence of events, that is $A_1 \supset A_2 \supset A_3, \dots$, then*

$$\mathbf{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n).$$

Remark 1.8. The last proposition is a not more than an application of the continuity of the probability function.

Proposition 1.7 (Continuity of the probability function). *Let $(A_n)_{n=1}^{\infty}$ be a sequence of events in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. If the limit $\lim_{n \rightarrow \infty} A_n$ exists, then $\lim_{n \rightarrow \infty} A_n \in \mathcal{F}$, and*

$$\mathbf{P} \left(\lim_{n \rightarrow \infty} A_n \right) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n).$$

Proof. We will prove this theorem for the case where $(A_n)_{n=1}^{\infty}$ is increasing. Define the following sequence of events:

$$\begin{aligned} B_1 &:= A_1, \\ B_n &:= A_n \setminus A_{n-1}. \end{aligned}$$

It is clear that:

- (1) The sets $(B_n)_{n=1}^{\infty}$ are disjoint.
- (2) For every $N \in \mathbb{N}$ we have:

$$\bigcup_{n=1}^N B_n = \bigcup_{n=1}^N A_n = A_N.$$

- (3) $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$.

We now have:

$$\begin{aligned} \mathbf{P} \left(\bigcup_{n=1}^{\infty} A_n \right) &= \mathbf{P} \left(\bigcup_{n=1}^{\infty} B_n \right) \\ &= \sum_{n=1}^{\infty} \mathbf{P}(B_n) \\ &= \lim_{N \rightarrow \infty} \sum_{n=1}^N \mathbf{P}(B_n) \\ &= \lim_{N \rightarrow \infty} \mathbf{P} \left(\bigcup_{n=1}^N B_n \right) \\ &= \lim_{N \rightarrow \infty} \mathbf{P}(A_N). \end{aligned}$$

□

2 Conditional probability

2.1 Conditional probability

Definition 2.1 (Conditional probability). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $A, B \in \mathcal{F}$, such that $\mathbf{P}(B) > 0$. We define the conditional probability of A “given that B already happened” by

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

Remark 2.1. The intuition behind this definition should be clear. We calculate the probability of the event A “inside” event B . Notice that we can also use conditional probability to calculate the the probability of an intersection of two events.

Proposition 2.1. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $B \in \mathcal{F}$ be an event such that $\mathbf{P}(B) > 0$. Then, the map $A \mapsto \mathbf{P}(A | B)$ is a probability function.

Proof. The proof that the range of the function is $[0, 1]$ and that $(\Omega | B) = 1$ is clear from the definitions, so we will only prove sigma additivity. Let $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ be disjoint sets, then $(A_n \cap B)_{n=1}^{\infty} \subset \mathcal{F}$ are also disjoint sets and we have:

$$\begin{aligned} \mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n | B\right) &= \frac{\mathbf{P}\left(\left(\bigcup_{n=1}^{\infty} A_n\right) \cap B\right)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}\left(\bigcup_{n=1}^{\infty} (A_n \cap B)\right)}{\mathbf{P}(B)} \\ &= \sum_{n=1}^{\infty} \frac{\mathbf{P}(A_n \cap B)}{\mathbf{P}(B)} \\ &= \sum_{n=1}^{\infty} \mathbf{P}(A_n | B), \end{aligned}$$

which concludes the proof. □

2.2 The law of total probability

Definition 2.2 (Partition). Let Ω be any set, let $N \in \mathbb{N} \cup \{\infty\}$, and let $\{A_n\}_{n=1}^{\infty} \subset 2^{\Omega}$, so that $\{A_n\}_{n=1}^{\infty}$ are (pairwise) disjoint and $\bigcup_{n=1}^N A_n = \Omega$. Then $\{A_n\}_{n=1}^N$ is called a partition of Ω .

Proposition 2.2 (Law of total probability). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\{A_n\}_{n=1}^N \subset 2^{\mathcal{F}}$ be a partition of Ω . Then,

$$\mathbf{P}(B) = \sum_{n=1}^N \mathbf{P}(A_n) \mathbf{P}(B | A_n).$$

More generally, if $\{A_n \cap B\}_{n=1}^N \subset 2^{\mathcal{F}}$ is a partition of B , then the same holds:

$$\mathbf{P}(B) = \sum_{n=1}^N \mathbf{P}(A_n) \mathbf{P}(B | A_n).$$

Remark 2.2. The set $\{A_n\}_{n=1}^N$ is called a partition of Ω .

Proof. The idea behind the proof should be very intuitive. If we think about it from a set-theoretic point of view, then clearly $\{A_n \cap B\}_{n=1}^N$ is a partition of B , and then $\mathbf{P}(B) = \sum_{n=1}^N \mathbf{P}(A_n \cap B)$, but by the definition of conditional probability we get

$$\mathbf{P}(B) = \sum_{n=1}^N \mathbf{P}(B \mid A_n) \mathbf{P}(A_n).$$

If we write it down, we get the same result:

$$\begin{aligned} \mathbf{P}(B) &= \mathbf{P}(B \cap \Omega) \\ &= \mathbf{P}\left(B \cap \bigcup_{n=1}^N A_n\right) \\ &= \mathbf{P}\left(\bigcup_{n=1}^N (A_n \cap B)\right) \\ &= \sum_{n=1}^N \mathbf{P}(A_n \cap B) \\ &= \sum_{n=1}^N \mathbf{P}(A_n) \mathbf{P}(B \mid A_n). \end{aligned}$$

□

Remark 2.3. Equivalently we could get from $\mathbf{P}(B) = \sum_{n=1}^N \mathbf{P}(A_n \cap B)$ to

$$\mathbf{P}(B) = \sum_{n=1}^N \mathbf{P}(A_n \mid B) \mathbf{P}(B),$$

but this isn't really helpful because we don't know the value of $\mathbf{P}(B)$.

Exercise 2.1 (Pólya's urn, simplified). Let there be 1 white and 1 black ball in an urn. At each step, one ball is drawn uniformly at random from the urn, and its color observed; it is then returned in the urn, and an additional ball of the same color is added to the urn. What is the probability that there are k black balls in the urn after the n th step?

Solution. First denote:

$$\begin{aligned} A_{n,k} &= \{\text{there are } k \text{ black balls after the } n\text{-th step.}\} \\ p_{n,k} &= \mathbf{P}(A_{n,k}). \end{aligned}$$

In order for there to be k black balls after the n -th step, there must either have been $k-1$ or k black balls in the $n-1$ -th step. Note that $\{A_{n-1,k} \cap A_{n,k}, A_{n-1,k-1} \cap A_{n,k}\}$ is a partition of $A_{n,k}$, so by the law of total probability we get

$$\mathbf{P}(A_{n,k}) = \mathbf{P}(A_{n-1,k-1}) \mathbf{P}(A_{n,k} \mid A_{n-1,k-1}) + \mathbf{P}(A_{n-1,k}) \mathbf{P}(A_{n,k} \mid A_{n-1,k}).$$

This implies that

$$p_{n,k} = \frac{k-1}{n+1} p_{n-1,k-1} + \frac{n+1-k}{n+1} p_{n-1,k}.$$

Coupled with the fact that $p_{0,1} = 1$ we can verify that the only solution under these conditions is $p_{n,k} = \frac{1}{n+1}$. In general, these problems are very hard to solve.

2.3 Bayes' theorem

Another useful trick is Bayes' theorem. In its simplified version it states that,

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

The proof in this case is very simple, however, there is a more general version to this simple formula.

Theorem 2.3 (Bayes' theorem). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let $\{A_n\}_{n=1}^N \subset 2^{\mathcal{F}}$ be a partition of Ω . Then,*

$$\mathbf{P}(A_i | B) = \frac{\mathbf{P}(B | A_i)\mathbf{P}(A_i)}{\sum_{n=1}^N \mathbf{P}(A_n)\mathbf{P}(B | A_n)},$$

for all $1 \leq i \leq N$ (remember N may be ∞).

Remark 2.4 (Intuition for Bayes' theorem). In the simple version of the theorem, notice we can rewrite the equation in a less standard form like so:

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

The left hand side is the probability that B happened, and also A happened given B happened, so in total, this is the probability that both A and B happened. From similar reasoning we get that the right hand side is the probability that both B and A happened. Obviously an equality must hold, and by understanding the statement in this way we can derive the equality by ourselves. Note that the more general equality is just the simple version with $A = A_i$ and $B = B$ and using the law of total probability.

Definition 2.3 (False positives etc.). Let Z be a binary condition a person can either have or not have (such as a disease), and suppose we have an artificial test to determine whether a person has condition Z or not, which returns either a positive, or a negative result. For simplicity, we assume condition Z is a disease. Then if we denote

$$\begin{aligned}\Omega &= \{\text{the person is sick, the person is healthy}\} \times \{\text{test positive, test negative}\} \\ A &= \{\text{the person is healthy}\} \\ B &= \{\text{the test result is positive}\}\end{aligned}$$

we can define the following:

True positive rate: the probability the test is positive given the person is sick: $\mathbf{P}(B | A^c)$

False negative rate: the probability the test is negative given the person is sick: $\mathbf{P}(B^c | A^c)$

True negative rate: the probability the test is negative given the person is healthy: $\mathbf{P}(B^c | A)$

False positive rate: the probability the test is positive given the person is healthy: $\mathbf{P}(B | A)$

Exercise 2.2. Suppose we have a test for checking whether a person is infected with the terrible the terrible disease called “the cooties”. Our test has a true positive rate of 0.98, and a false positive rate of 0.01. Suppose that 0.1% of the population has the cooties. What is the probability that a person is sick, given that the test result is positive?

Solution. Denote

$$A = \{\text{the person is healthy}\} \quad \text{and} \quad B = \{\text{the result is positive}\}.$$

From Bayes' theorem we have:

$$\mathbf{P}(A | B) = \frac{\mathbf{P}(B | A)\mathbf{P}(A)}{\mathbf{P}(B)} = \frac{0.01 \cdot 0.999}{\mathbf{P}(B)}.$$

From the law of total probability we have

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A)\mathbf{P}(B | A) + \mathbf{P}(A^c)\mathbf{P}(B | A^c) \\ &= 0.01 \cdot 0.999 + 0.98 \cdot 0.001 = 0.01097.\end{aligned}$$

And thus,

$$\mathbf{P}(A | B) = \frac{0.01 \cdot 0.999}{0.01097} \approx 0.91$$

which implies that $\mathbf{P}(A^c | B) \approx 0.09$.

Exercise 2.3. Suppose you are going to lecture about Riemann surfaces, so that the probability of n students arriving to your class is $p_n = \frac{2}{(n+2)(n+3)}$. Then you ask a question, and all students raise their hands (a rare occasion indeed), but since you are really passionate about Riemann surfaces, there's an equal chance you could answer your own question, so that the probability of any living mathematician in class to answer your question is $\frac{1}{n+1}$. In these conditions

- (1) What is the probability that you answered your own question?
- (2) What is the probability that exactly 3 people arrived to your lecture, given that you answered your own question?

Solution. First of all we notice that

$$\sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \frac{2}{(n+2)(n+3)} = \sum_{n=0}^{\infty} \left\{ \frac{2}{n+2} - \frac{2}{n+3} \right\} = 1 - 0 = 1,$$

so we can define a probability space under the conditions in the exercise.

- (1) Denote

$$\begin{aligned}A &= \{\text{you answered your own question}\}, \\ B_n &= \{\text{exactly } n \text{ people arrived to the lecture}\}.\end{aligned}$$

Now since $\{B_n\}_{n=0}^{\infty}$ is a partition of Ω , by the law of total probability

$$\begin{aligned}\mathbf{P}(A) &= \sum_{n=0}^{\infty} \mathbf{P}(A | B_n)\mathbf{P}(B_n) = \sum_{n=0}^{\infty} \frac{1}{n+1} \cdot \frac{2}{(n+2)(n+3)} \\ &= \sum_{n=0}^{\infty} \left(\frac{1}{n+1} - \frac{2}{n+2} + \frac{1}{n+3} \right) = 1 - \frac{1}{2} = \frac{1}{2}.\end{aligned}$$

- (2) From Bayes' theorem we get

$$\mathbf{P}(B_3 | A) = \frac{\mathbf{P}(A | B_3)\mathbf{P}(B_3)}{\mathbf{P}(A)} = \frac{\frac{1}{4} \cdot \frac{2}{5 \cdot 6}}{\frac{1}{2}} = \frac{1}{30},$$

which completes the proof.

3 Independence and repeating experiments

3.1 Independence of events

Intuitively, when we say that the event A is independent from B , we mean that

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

Thus we can use the definition of conditional probability in order to try to formalize the notion of independence of events. However, this definition only works when $\mathbf{P}(B) > 0$, and also it is not immediate from the definition that if A is independent from B , then the converse is also true (which is something we want, since independency is symmetrical). Thus, we will define independence in a similar, better way.

Definition 3.1 (Independence of two events). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, let $A, B \in \mathcal{F}$ be two events. We say that A and B are independent, if

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

Remark 3.1. Note that Definition 3.1 is a generalization of the intuitive definition $\mathbf{P}(A \mid B) = \mathbf{P}(A)$. This means that if $\mathbf{P}(A \mid B) = \mathbf{P}(A)$ holds, in particular A and B are independent.

Example 3.1. Recall Exercise 2.3. We want to determine for which n the events A and B_n are independent. In this case, we don't need to both to compute $\mathbf{P}(A \cap B_n)$, we just notice that $\mathbf{P}(A \mid B_n) = \mathbf{P}(A)$, or in numbers $\frac{1}{n+1} = \frac{1}{2}$ if and only if $n = 1$, and that is the answer.

Proposition 3.1. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, let $A \in \mathcal{F}$. The following conditions are equivalent:

- (1) For each $B \in \mathcal{F}$ the events A and B are independent.
- (2) A is independent of itself.
- (3) $\mathbf{P}(A) \in \{0, 1\}$.

Proof. Immediate from the definitions. □

Definition 3.2 (Independence of events). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, let $\{A_n\}_{n=1}^N$ be a finite sequence of events. We say that $\{A_n\}_{n=1}^N$ are independent if for each $\emptyset \neq K \subset \{1, 2, \dots, N\}$ we have

$$\mathbf{P}\left(\bigcap_{n \in K} A_n\right) = \prod_{n \in K} \mathbf{P}(A_n).$$

Definition 3.3 (Pairwise independent events). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\{A_n\}_{n=1}^N$ be a finite sequence of events. We say that $\{A_n\}_{n=1}^N$ are pairwise independent if for each $1 \leq i < j \leq N$ we have

$$\mathbf{P}(A_i \cap A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j).$$

Definition 3.4 (Independence of infinite events). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\{A_n\}_{n=1}^\infty$ be an infinite sequence of events. We say that $\{A_n\}_{n=1}^\infty$ are (pairwise) independent if each finite subset of them is (pairwise) independent.

Proposition 3.2. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, let $\{A_n\}_{n=1}^\infty$ be an infinite sequence of (pairwise) independent events. Then define a new sequence $\{\tilde{A}_n\}_{n=1}^\infty$ such that $\tilde{A}_n = A_n$ or $\tilde{A}_n = A_n^c$ for each $n \in \mathbb{N}$. Then each choice of such $\{\tilde{A}_n\}_{n=1}^\infty$ is (pairwise) independent.

Proof. Using induction on the number of indices where we chose $\tilde{A}_n = A_n^c$. □

Remark 3.2. Now we have the tools to define probability spaces for repeating experiments. We always assume that the experiment is repeated in exactly the same way, and that the results of each experiment are independent.

Definition 3.5 (Probability space of repeating trials, finite case). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space for a certain experiment. We define the probability space for repeating the experiment $N < \infty$ times as follows:

$$\begin{aligned}\Omega_N &= \Omega^N. \\ \mathcal{F}_N &= \sigma \left(\{A_1 \times \cdots \times A_n \mid A_1, \dots, A_n \in \mathcal{F}\} \right). \\ \mathbf{P}_N(A_1 \times A_2 \times \cdots \times A_N) &= \prod_{i=1}^N \mathbf{P}(A_i).\end{aligned}$$

The fact that $\mathbf{P}_N$ is a probability measure follows from Theorem 1.4. This probability measure is called the product measure, and $\mathcal{F}_N$ is called the product σ -algebra.

Definition 3.6 (Probability space of repeating trials, infinite case). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space for a certain experiment. We define the probability space for repeating the experiment $N = \infty$ times as follows:

$$\begin{aligned}\Omega_N &= \Omega^{\mathbb{N}}. \\ \mathcal{F}_N &= \sigma \left(\left\{ \prod_{i=1}^{\infty} A_i \mid \forall i \geq 1 (A_i \in \mathcal{F}) \text{ and only for finitely many } i\text{'s } A_i \neq \Omega \right\} \right). \\ \mathbf{P}_N \left(\prod_{i=1}^N A_i \right) &= \prod_{i=1}^N \mathbf{P}(A_i).\end{aligned}$$

Remark 3.3. Note that in both cases we did not actually define $\mathbf{P}_N$ for all events in $\mathcal{F}_N$. The extension to the rest of the events is by an application of Theorem 1.4.

Remark 3.4. Note that the definition for the case $N = \infty$ generalizes the definition for the $N < \infty$ case. We don't actually need both definitions, the latter covers both cases. We only defined the cases separately for clarity.

3.2 Bernoulli Trials

Example 3.2 (Bernoulli trial). A Bernoulli trial, is a random experiment with only two possible outcomes, “success” and “failure”, in which the probability of success is the same every time the experiment is conducted. The probability space that models these kind of experiments is given by:

$$\begin{aligned}\Omega &= \{0, 1\}. \\ \mathcal{F} &= \{\emptyset, \{0\}, \{1\}, \Omega\}. \\ \mathbf{P}(\omega) &= \begin{cases} p, & \omega = 1 \\ 1 - p, & \omega = 0 \end{cases}\end{aligned}$$

which models repeating independent experiment with two results. These kind of experiments are also called *Bernoulli trials*.

Let $N \in \mathbb{N}$. We want to calculate the probability that the experiment repeated N times ends in k successes. We set:

$$A_k = \{k \text{ successes}\}.$$

$$H_i = \{\omega_i = 1\} = \{\text{the } i\text{th experiment was a success}\},$$

for all $1 \leq i \leq N$. We now notice that if $\omega = (\omega_1, \dots, \omega_N) \in A_k$ then

$$\{\omega\} = \bigcap_{i=1}^N \tilde{H}_i,$$

where

$$\tilde{H}_i = \begin{cases} H_i & \omega_i = 1, \\ H_i^c & \omega_i = 0. \end{cases}$$

Since the results of each experiment are independent we can use Theorem 3.2 to get

$$\mathbf{P}_N(\{w\}) = \mathbf{P}_N\left(\bigcap_{i=1}^N \tilde{H}_i\right) = \prod_{i=1}^N \mathbf{P}_N(\tilde{H}_i) = p^k(1-p)^{N-k}$$

Finally, we conclude that

$$P_N(A_k) = |A_k| p^k(1-p)^{N-k} = \binom{N}{k} p^k(1-p)^{N-k}$$

Also, because we know that $\{A_k\}_{k=0}^N$ are a partition of Ω_N , we can verify our answer by making sure $\sum_{k=0}^N \mathbf{P}_N(A_k) = 1$. Indeed,

$$\sum_{k=0}^N \mathbf{P}_N(A_k) = \sum_{k=0}^N \binom{N}{k} p^k(1-p)^{N-k} = (p + (1-p))^N = 1.$$

3.3 Random walks

Example 3.3 (Random walks). Let there be a cute cat on the $\mathbb{Z}$ number line. We know that each minute, the cat could move one step to the right with a probability of p , or one step to the left with probability $1-p$. We may wonder what are the chances the cat would be on the number 4 after 8 minutes.

It's not hard to see that this experiment is a Bernoulli trial, and indeed if we denote the number of step the cat made to the right after 8 minutes by A_k , and denote the product probability space by $(\Omega_8, \mathcal{F}_8, \mathbf{P}_8)$, we would get that:

$$\mathbf{P}_8(A_k) = \binom{8}{k} p^k(1-p)^{8-k}.$$

The position of the cat after k steps to the right is $k - (8-k) = 2k - 8$. Since we wonder about the probability of it being on the number 4, we need calculate the probability of A_6 . The result is

$$\mathbf{P}_8(A_6) = \binom{8}{6} p^6(1-p)^{8-6}.$$

3.4 Infinite coin flips

Example 3.4 (Infinite coin flips). We now consider the probability space $(\Omega_\infty, \mathcal{F}_\infty, \mathbf{P}_\infty)$ corresponding to the product space of infinite Bernoulli experiments. We want to calculate the probability that the first success was in the n th experiment. Denote:

$$H_i = \{\text{the } i\text{th experiment resulted in success}\}$$

$$R_n = \{\text{first success was in the } n\text{th experiment}\} = \left(\bigcap_{i=1}^{n-1} H_i^c \right) \cap H_n.$$

Since all the Bernoulli experiments are independent, from Theorem 3.2 we get

$$\mathbf{P}_\infty(R_n) = \mathbf{P}_\infty \left(\left(\bigcap_{i=1}^{n-1} H_i^c \right) \cap H_n \right) = \left(\prod_{i=1}^{n-1} \mathbf{P}_\infty(H_i^c) \right) \cdot \mathbf{P}_\infty(H_n) = (1-p)^{n-1}p.$$

All the events $\{R_n\}_{n=1}^\infty$ are all disjoint, and also

$$\bigcup_{n=1}^\infty R_n = \Omega_\infty \setminus \{(0, 0, 0, \dots)\},$$

Moreover, for $p > 0$, we may assume that $\mathbf{P}_\infty(0, 0, 0, \dots) = 0$. This allows us to compute $\mathbf{P}_\infty(0, 0, 0, \dots) = 0$ when $p > 0$, as follows:

$$\begin{aligned} \mathbf{P}_\infty(\{0, 0, 0, \dots\}) &= 1 - \mathbf{P}_\infty(\Omega_\infty \setminus \{0, 0, 0, \dots\}) = 1 - \mathbf{P}_\infty\left(\bigcup_{n=1}^\infty R_n\right) \\ &= 1 - \sum_{n=1}^\infty \mathbf{P}_\infty(R_n) = 1 - p \sum_{n=1}^\infty (1-p)^{n-1} \\ &= \begin{cases} 0 & p \in (0, 1), \\ 0 & p = 1, \end{cases} \end{aligned}$$

where the last step follows from the formula for an infinite geometric series. Finally, when $p = 0$ it is clear that $\mathbf{P}_\infty(0, 0, 0, \dots) = 1$.

4 Random variables

In this section we will introduce the notion of random variables. In a general experiment take a certain result $\omega \in \Omega$. In real life, we usually want to know multiple things about each result, and maybe categorize results based on a certain *variables*. For example when we roll two dice, a result $\omega = (2, 4) \in \Omega$ has the property that the sum of the dice is 3. It also have the property that the second dice result is two times the result of the first dice. However while the sum of results of $\eta = (4, 2) \in \Omega$ is 6, here the second dice's result is half the result of the first dice.

In order to do probability on all of these properties we need some map, which we call a *random variables*, that sends each $\omega \in \Omega$, to a value in $\{0, 1\}$ or $\{2, \dots, 12\}$. In general the results of random variables are real numbers, so we will define their range to be $\mathbb{R}$.

It is also worth noting that in order for the probability function to be defined on the per-images of random variables, we must define the preimage of some basic sets in $\mathbb{R}$ to be events in the σ -algebra.

4.1 Random variables

Definition 4.1 (Random variable). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A function $X: \Omega \rightarrow \mathbb{R}$ is said to be a random variable if for any open interval $(a, b) \subset \mathbb{R}$ we have

$$X^{-1}((a, b)) = \{\omega \in \Omega \mid X(\omega) \in (a, b)\} \in \mathcal{F}.$$

Remark 4.1. From now on we denote

$$\{X \in A\} = \{\omega \in \Omega \mid X(\omega) \in A\}$$

Here are a couple of things we should notice:

- (1) If we have that Ω is countable then as we know $\mathcal{F} = 2^\Omega$ and thus every function $X: \Omega \rightarrow \mathbb{R}$ is a random variable.
- (2) If we denote

$$\mathcal{G}_X := \{D \subset \mathbb{R} \mid \{X \in D\} \in \mathcal{F}\},$$

we can notice that $\mathbb{R} \in \mathcal{G}_X$, and that $\mathcal{G}_X$ is closed under countable unions, and complements. Thus, it is a σ -algebra of $\mathbb{R}$. We also have that it contains all the open intervals in $\mathbb{R}$ so we get that $\mathfrak{B} := \mathfrak{B}(\mathbb{R}) \subset \mathcal{G}_X$.

Definition 4.2 (Distribution of a random variable). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and $X: \Omega \rightarrow \mathbb{R}$ a random variable. The distribution of X , denoted $\mathbf{P}_X$ is the function $\mathbf{P}_X: \mathfrak{B} \rightarrow [0, 1]$ defined as such:

$$\mathbf{P}_X(A) = \mathbf{P}(X \in A).$$

Remark 4.2. Note that $\mathbf{P}_X$ is well defined since $\mathfrak{B} \subset \mathcal{G}_X$.

Remark 4.3. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Then the space $(\mathbb{R}, \mathfrak{B}, \mathbf{P}_X)$ is also a probability space.

Remark 4.4. Random variable X gives us data about experiments. Thus, for events $N \in \mathcal{F}$ such that $\mathbf{P}(N) = 0$, the result of $X(N)$ is meaningless. This justifies not defining random variables on events with probability 0.

Exercise 4.1. There are 20 balls in a vase, numbered from 1 – 20. Three balls are taken out of the vase with uniform probability. What is the probability that one of the balls is numbered 17 or above?

Solution. First we define our sample space

$$\Omega = \{(i, j, k) \mid 1 \leq i \leq j \leq k \leq 20\}.$$

Since the balls are taken out uniformly we have

$$\mathbf{P}((i, j, k)) = \binom{20}{3}^{-1}.$$

We now define the relevant random variable $X: \Omega \rightarrow \mathbb{R}$ by

$$X((i, j, k)) = k,$$

and we notice that we want to calculate $\mathbf{P}(X \in \{17, 18, 19, 20\})$. Since $\mathbf{P}_X$ is a probability function and the events $\mathbf{P}_X(\{m\})_{m=17}^{20}$ are disjoint we get:

$$\begin{aligned} \mathbf{P}(X \in \{17, 18, 19, 20\}) &= \mathbf{P}_X(\{17, 18, 19, 20\}) \\ &= \mathbf{P}_X(\{17\}) + \mathbf{P}_X(\{18\}) + \mathbf{P}_X(\{19\}) + \mathbf{P}_X(\{20\}) + \end{aligned}$$

Since $\{X \in \{k\}\}$ are exactly the events we pick 2 balls with value less than k , we have that $\mathbf{P}_X(\{k\}) = \binom{k-1}{2} / \binom{20}{3}$. Thus, we can conclude that

$$\mathbf{P}(X \in \{17, 18, 19, 20\}) = \binom{20}{3}^{-1} \left[\binom{16}{2} + \binom{17}{2} + \binom{18}{2} + \binom{19}{2} \right] \approx 0.508.$$

4.2 Coupon collector

Example 4.1 (Coupon collector). Suppose there are N types of coupons in a certain game, and we want to collect them all. Thus, each day we buy a new random coupon, with uniform probability to be any coupon in the game (including ones we already have). The sample space is we define is $\Omega = \{1, 2, \dots, N\}^{\mathbb{N}}$, which denotes which coupon we got on each day. We define a random variable $T: \Omega \rightarrow \mathbb{R}$ by

$$T(\omega) = \inf \left\{ k \geq 1 : |\{\omega_1, \dots, \omega_k\}| = N \right\}.$$

Notice that T is undefined for $\omega \in \Omega$ that doesn't include all the coupons, so we need to show that the probability of such events is zero. It would be easier to do it later, so we will assume that for now. For $t \geq N$ and $1 \leq j \leq N$ we define

$$A_t(j) = \{\text{we didn't find the coupon of type } j \text{ up to day } t\}.$$

Notice that:

$$\{T > t\} = \bigcup_{j=1}^N A_t(j).$$

Thus,

$$\mathbf{P}_T((t, \infty)) = \mathbf{P}_T(\{t+1, t+2, \dots\}) = \mathbf{P}\left(\bigcup_{j=1}^N A_t(j)\right)$$

The events $A_t(j)$ are not disjoint but using the inclusion-exclusion principle we get

$$\mathbf{P}\left(\bigcup_{j=1}^N A_t(j)\right) = \sum_{t=1}^N (-1)^{i+1} \sum_{\substack{J \subset \{1, 2, \dots, N\} \\ |J|=i}} \mathbf{P}\left(\bigcap_{j \in J} A_t(j)\right).$$

We can notice that

$$\mathbf{P} \left(\bigcap_{j \in J} A_t(j) \right) = \left(\frac{N - |J|}{N} \right)^t.$$

Summing up the results so far gives

$$\mathbf{P}_T(\{t+1, t+2, \dots\}) = \sum_{\substack{J \subset \{1, 2, \dots, N\} \\ |J|=i}} \left(\frac{N - |J|}{N} \right)^t = \sum_{i=1}^N (-1)^{i+1} \binom{N}{i} \left(\frac{N-i}{N} \right)^t.$$

Because the image of T is $\mathbb{N}$ we can write

$$\begin{aligned} \mathbf{P}_T(\{t\}) &= \mathbf{P}_T(\{t, t+1, \dots\}) - \mathbf{P}_T(\{t+1, t+2, \dots\}) \\ &= \sum_{i=1}^N (-1)^{i+1} \binom{N}{i} \left(\frac{N-i}{N} \right)^{t-1} - \sum_{i=1}^N (-1)^{i+1} \binom{N}{i} \left(\frac{N-i}{N} \right)^t \\ &= \sum_{i=1}^N (-1)^{i+1} \binom{N}{i} \left(\frac{N-i}{N} \right)^t \left(1 - \frac{N-i}{N} \right). \end{aligned}$$

Now all that's left to show is that the probability of events where we didn't get all the coupons is zero. For $1 \leq j \leq N$ we define the event $A(j) = \{\text{we didn't find the coupon of type } j\}$ and notice that:

$$A(j) = \bigcap_{t=1}^{\infty} A_t(j).$$

Since $A_t(j)$ is a decreasing sequence of events (in t), it is clear that $A(j) \in \mathcal{F}$. Since probability functions are continuous we have that

$$\mathbf{P}(A(j)) = \mathbf{P} \left(\bigcap_{t=1}^{\infty} A_t(j) \right) = \lim_{t \rightarrow \infty} \mathbf{P}(A_t(j)) = \lim_{t \rightarrow \infty} \left(\frac{N-1}{N} \right)^t = 0,$$

therefore,

$$\mathbf{P}(\text{we didn't find one of the coupons}) = \mathbf{P} \left(\bigcup_{j=1}^N A(j) \right) \leq \sum_{j=1}^N \mathbf{P}(A(j)) = \sum_{j=1}^N 0 = 0.$$

4.3 Cumulative distribution function

Definition 4.3 (Support). The support of a random variable $X: \Omega \rightarrow \mathbb{R}$ is the set of all $a \in \mathbb{R}$ such that for all $\varepsilon > 0$ we have $\mathbf{P}_X((a - \varepsilon, a + \varepsilon)) > 0$. For a general function $f: A \rightarrow \mathbb{R}$ such that A is a topological space we have

$$\text{supp}(f) := \text{cl}_A(\{x \in A : f(x) \neq 0\}) = \overline{f^{-1}(\{0\}^c)}.$$

We usually denote the support of X by $\text{supp}(\mathbf{P}_X)$. Note that this coincides with the general definition of support only when we define $\mathbf{P}_X: \mathbb{R} \rightarrow [0, 1]$ by $\mathbf{P}_X(a) = \mathbf{P}_X(\{a\})$.

Definition 4.4 (Cumulative distribution function). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. The cumulative distribution function (CDF) of X , denoted F_X is a function $F_X: \mathbb{R} \rightarrow [0, 1]$ defined as such

$$F_X(a) := \mathbf{P}_X((-\infty, a]) \equiv \mathbf{P}(X \leq a).$$

Remark 4.5. The cumulative distribution functions F_X is sometimes simply called the distribution function of X .

Proposition 4.1. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. Then the cumulative distribution function F_X satisfies the following properties:

- (1) F_X is a monotonically increasing function.
- (2) $\lim_{a \rightarrow \infty} F_X(a) = 1$.
- (3) $\lim_{a \rightarrow -\infty} F_X(a) = 0$.
- (4) F_X is right-continuous.

Proof. The proofs of 1 – 3 are derived from basic properties of probability functions, and that they are continuous, so we will only prove the last statement. Let $a \in \mathbb{R}$. Since F_X is monotonically increasing, it suffices to show the continuity for a single sequence, for example $a_n = a + \frac{1}{n}$. We see that

$$\begin{aligned} \lim_{n \rightarrow \infty} F_X\left(a + \frac{1}{n}\right) &= \lim_{n \rightarrow \infty} \mathbf{P}_X((-\infty, a + 1/n]) = \mathbf{P}_X\left(\bigcap_{n=1}^{\infty} (-\infty, a + 1/n]\right) \\ &= \mathbf{P}_X((-\infty, a]) = F_X(a), \end{aligned}$$

where the second equality follows from the continuity of $\mathbf{P}_X$. □

Remark 4.6. A cumulative distribution function is not always left-continuous. Let $x \in \mathbb{R}$, we have

$$\begin{aligned} \lim_{h \rightarrow 0^+} F_X(x - h) &= \lim_{h \rightarrow 0^+} \mathbf{P}_X((-\infty, x - h]) = \mathbf{P}_X\left(\bigcup_{n=1}^{\infty} (-\infty, x - h]\right) \\ &= \mathbf{P}_X((-\infty, x)) = F_X(x) - \mathbf{P}_X(x), \end{aligned}$$

and for points x such that $\mathbf{P}_X(x) \neq 0$, we get that F_X is discontinuous from the left.

Remark 4.7 (Intuition for one-sided continuity). Recall that $F_X(a)$ computes the probability of the set $(-\infty, a]$, and so if we go from the right, the limit of $(-\infty, a + h]$ as $h \rightarrow 0^+$ is $(-\infty, a]$ which shows that right-continuity makes sense. However, if we go from the other we have $(-\infty, a - h]$ where $h \rightarrow 0^+$ is converging to the set $(-\infty, a)$, which is different from $(-\infty, a]$.

Theorem 4.2. Every cumulative distribution function corresponds to a unique distribution. In other words, if $(\Omega, \mathcal{F}, \mathbf{P})$ and $(\Omega', \mathcal{F}', \mathbf{P}')$ are two probability spaces, and $X: \Omega \rightarrow \mathbb{R}$, $Y: \Omega' \rightarrow \mathbb{R}$ are two random variables, then

$$F_X = F_Y \iff \mathbf{P}_X = \mathbf{P}_Y.$$

Proof. Outside the scope of the course. □

Remark 4.8. Theorem 4.2 is the reason both $\mathbf{P}_X$ and F_X are distribution functions. However, since F_X is cumulative, it is called the *cumulative* distribution function, while $\mathbf{P}_X$ is simply called the distribution function.

Theorem 4.3. Let $F: \mathbb{R} \rightarrow [0, 1]$ be a function that satisfies all 4 basic properties described in Theorem 4.1. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a random variable $X: \Omega \rightarrow \mathbb{R}$ such that $F_X = F$.

Proof. Outside the scope of the course. $\square$

Remark 4.9. From two last theorems we can conclude that there exists a bijection between random variables' distributions and functions that satisfy the 4 properties described in Theorem 4.1.

Definition 4.5 (Discrete random variable). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. We say that X is a discrete random variable if $\text{supp}(\mathbf{P}_X)$ is a countable set.

Remark 4.10. Notice that if $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space, and Ω is countable, then all random variables $\Omega \rightarrow \mathbb{R}$ are discrete.

Remark 4.11 (Discrete variables and step functions). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a discrete random variable. Denote $\text{supp}(\mathbf{P}_X) = \{x_i\}_{i=1}^{\infty}$ and $p_i := \mathbf{P}(\{x_i\})$. We now have that

$$\mathbf{P}_X((a, b)) = \sum_{i: x_i \in (a, b)} p_i,$$

and therefore,

$$F_X(a) = \mathbf{P}_X((-\infty, a)) = \sum_{i: x_i \leq a} p_i.$$

This means that the CDF of a discrete random variable X is an a step function. In fact, a random variable X is discrete if and only if F_X is a step function.

4.4 Binomial distribution

Definition 4.6 (Binomial distribution). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. We say that X is a binomial random variable with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ if its range is $\{0, 1, 2, \dots, n\}$ and its distribution satisfies

$$\mathbf{P}_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \forall k \in \{0, 1, 2, \dots, n\}.$$

Alternatively, X is a binomial random variable with parameters p and n if

$$F_X(a) = \sum_{\substack{k \leq a \\ k \in \{0, 1, 2, \dots, n\}}} \binom{n}{k} p^k (1-p)^{n-k}, \quad \forall a \in \mathbb{R}.$$

When X is a binomial random variable, we denote $X \sim \text{Bin}(n, p)$.

Remark 4.12. We saw in Example 3.2 that the random variable $X: \Omega_N \rightarrow \mathbb{R}$ defined by $X(\omega) = \#\text{wins}$ is a binomial random variable with parameters n and p , so we can write $X \sim \text{Bin}(N, p)$.

Example 4.2. It is said that Hercules had a 90% chance to complete each one of his 12 deadly labours. We all know that he eventually completed all of them, and gained immortality. But how easier would it be, if he was allowed to fail once or twice? In this case it is clear that the random variable $X(\omega) = \#\text{fails}$ satisfies $X \sim \text{Bin}(12, 0.1)$.

$$\begin{aligned} \mathbf{P}_X(\{1, 2\}) - \mathbf{P}_X(\{0\}) &= \mathbf{P}_X(\{1\}) + \mathbf{P}_X(\{2\}) - \mathbf{P}_X(\{0\}) \\ &= \sum_{k \in \{1, 2\}} \binom{12}{k} (0.1)^k (1 - (0.1))^{12-k} - \binom{12}{0} (0.1)^0 (1 - (0.1))^{12-0} \\ &\approx 0.325 \end{aligned}$$

Example 4.3. Suppose we flip a fair coin 2000 times. You may be able to convince your friend it's safe to bet on the chances it falls on heads exactly 1000 times. After all, it's a fair coin, so flipping it a large number of times means it's safe to assume it fell on head 50% of the time right? Let's find out. Let $X(\omega) = \# \text{heads}$ so that $X \sim \text{Bin}(2000, 0.5)$. We get

$$\mathbf{P}_X(1000) = \binom{2000}{1000} \left(\frac{1}{2}\right)^{1000} \left(\frac{1}{2}\right)^{1000} = \frac{(2000)!}{(1000!)^2 2^{2000}}.$$

Using Stirling's approximation

$$\lim_{n \rightarrow \infty} \frac{n!}{\sqrt{2\pi n} n^{n+1/2} e^{-n}} = 1,$$

we get that

$$\mathbf{P}_X(1000) \approx \frac{\sqrt{2\pi}(2000)^{2000+1/2} e^{-2000}}{(\sqrt{2\pi}1000^{1000+1/2} e^{-1000})^2 2^{2000}} = \frac{1}{\sqrt{\pi}1000}.$$

Turns out it wasn't such a good idea...

4.5 Geometric distribution

Definition 4.7 (Geometric distribution). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. We say that X is a geometric random variable with parameter $p \in [0, 1]$ if its range is $\mathbb{N}$ and its distribution satisfies

$$\mathbf{P}_X(k) = (1-p)^{k-1}p, \quad \forall k \in \mathbb{N}.$$

Alternatively, X is a geometric random variable with parameter p if

$$F_X(a) = \sum_{\substack{k \leq a \\ k \in \mathbb{N}}} (1-p)^{k-1}p, \quad \forall a \in \mathbb{R}.$$

In this case we denote $X \sim \text{Geo}(p)$.

Example 4.4. Suppose we had a Bernoulli experiment repeated indefinitely, or until we get a success. Let p denote the probability of success. The random variable that calculates the first repetition in which the experiment succeeded follows $\text{Geo}(p)$.

Definition 4.8 (Poisson distribution). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. We say that X is a Poisson random variable with parameter $\lambda > 0$ if its range is $\mathbb{N}$ and its distribution satisfies

$$\mathbf{P}_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad \forall k \in \mathbb{N}.$$

Alternatively, X is a Poisson random variable with parameter λ if

$$F_X(a) = \sum_{\substack{k \leq a \\ k \in \mathbb{N}}} e^{-\lambda} \frac{\lambda^k}{k!}, \quad \forall a \in \mathbb{R}.$$

In this case we denote $X \sim \text{Poi}(\lambda)$.

Remark 4.13. The Poisson distribution often comes up when we try to model certain physical events. Suppose that a certain event happens at a rate of λ . We can try to model this by thinking of repeating a lot of trials, say n of them, and in each there is a probability λ/n of succeeding. As we take $n \rightarrow \infty$ the more gets more and more accurate, and at the limit, it will follow Poisson distribution.

Theorem 4.4 (Poisson approximation to binomial). *Suppose that $X \sim \text{Bin}(n, k)$ and we have $np = \lambda$. Then*

$$\mathbf{P}_X(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \xrightarrow{n \rightarrow \infty} \frac{\lambda^k}{k!} e^{-\lambda}.$$

Proof.

$$\begin{aligned} \mathbf{P}_X(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n(n-1) \cdots (n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \cdot \underbrace{\left(\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n}\right)}_{(*)} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

As we take the limit $n \rightarrow \infty$ we have $(*) \rightarrow 1$ and thus

$$\lim_{n \rightarrow \infty} \mathbf{P}_X(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

□

4.6 Divergence of the sum of the reciprocals of the primes

Before moving on to continuous random variables, we will prove a theorem from a very unexpected field.

Theorem 4.5 (Divergence of the sum of the reciprocals of the primes). *We want to show that*

$$\sum_{p \text{ prime}} \frac{1}{p} = \frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots = \infty.$$

Definition 4.9 (Riemann zeta function). For any real number $s > 1$ the Riemann zeta function is defined as such

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}.$$

Lemma 4.6 (Euler's product formula).

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}}$$

where $\zeta(s)$ is visibly the Riemann zeta function, and s is a real number greater than 1.

Proof. Let $s > 1$ be a real number. Define a discrete random variable X , with range (support) $\{1, 2, \dots\}$, with the following distribution

$$\mathbf{P}_X(k) = \frac{k^{-s}}{\zeta(s)},$$

where $\zeta(s)$ is the Riemann zeta function. This assures us that $\mathbf{P}_X(\mathbb{N}) = 1$. For any natural number m set

$$A_m = \{k \in \mathbb{N} \mid k \text{ is divisible by } m\} = \{nm \mid n \in \mathbb{N}\}.$$

Since $\mathbf{P}_X$ is a probability function we get that for any $m \in \mathbb{N}$

$$\begin{aligned}\mathbf{P}_X(A_m) &= \sum_{k \in A_m} \mathbf{P}_X(k) = \sum_{n=1}^{\infty} \mathbf{P}_X(nm) = \frac{1}{\zeta(s)} \sum_{n=1}^{\infty} (nm)^{-s} \\ &= \frac{m^{-s}}{\zeta(s)} \sum_{n=1}^{\infty} n^{-s} = \frac{m^{-s}}{\zeta(s)} \zeta(s) = m^{-s}.\end{aligned}$$

We can notice that for any two distinct primes $p \neq q$ that the events A_q and A_p are independent. That is because $A_p \cap A_q = A_{pq}$, and then

$$\mathbf{P}_X(A_p \cap A_q) = \mathbf{P}_X(A_{pq}) = (pq)^{-s} = p^{-s} q^{-s} = \mathbf{P}_X(A_p) \mathbf{P}_X(A_q).$$

Similarly, we can show that for any finite number of primes $\{p_i\}_{i=1}^N$ that the events $\{A_{p_i}\}_{i=1}^N$ are independent, so by definition $\{A_p\}_{p \text{ prime}}$ are all independent. Finally, since every integer greater than 1 is divisible by at least one prime number

$$\bigcap_{p \text{ prime}} A_p^c = \{1\}.$$

Using these facts, we get that

$$\frac{1}{\zeta(s)} = \mathbf{P}_X(\{1\}) = \mathbf{P}_X\left(\bigcap_{p \text{ prime}} A_p^c\right) = \prod_{p \text{ prime}} \mathbf{P}_X(A_p^c) = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right),$$

which completes the proof. $\square$

Proof of Theorem 4.5. Because we already know from analysis that $\sum_{n \in \mathbb{N}} \frac{1}{n} = \infty$ we get that

$$\begin{aligned}\lim_{s \rightarrow 1^+} \exp\left(\sum_{p \text{ prime}} \ln\left(1 - \frac{1}{p^s}\right)\right) &= \lim_{s \rightarrow 1^+} \exp\left(\ln\left(\prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)\right)\right) \\ &= \lim_{s \rightarrow 1^+} \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right) \\ &= \frac{1}{\zeta(s)} \\ &= 0.\end{aligned}$$

Now from the properties of the exponent function we get

$$\lim_{s \rightarrow 1^+} \sum_{p \text{ prime}} \ln\left(1 - \frac{1}{p^s}\right) = -\infty.$$

Since for $0 < x < 0.6$ we have $\ln(1 - x) \geq -2x$ we get

$$\lim_{s \rightarrow 1^+} -2 \sum_{p \text{ prime}} \frac{1}{p^s} \leq \lim_{s \rightarrow 1^+} \sum_{p \text{ prime}} \ln\left(1 - \frac{1}{p^s}\right) = -\infty,$$

which proves that

$$\sum_{p \text{ prime}} \frac{1}{p} = \infty,$$

and completes the proof. $\square$

4.7 Continuous and absolutely continuous random variables

Definition 4.10 (Continuous random variable). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. We say that X is a continuous random variable if F_X is a continuous function.

Remark 4.14. Recall what we did in Remark 4.6. Note again that if X is a continuous random variable then for any $a \in \mathbb{R}$

$$\mathbf{P}(a) = \lim_{n \rightarrow \infty} ((a - 1/n, a]) = \lim_{n \rightarrow \infty} F_X(a) - F_X(a - 1/n) = 0.$$

Remark 4.15. Continuous random variables can get very weird, and in this course we will only discuss a very specific family of continuous random variables, called *absolutely continuous random variables*.

Definition 4.11 (Absolutely continuous random variable). A random variable X is said to be absolutely continuous if there exists an integrable function $f_X: \mathbb{R} \rightarrow [0, \infty)$ such that for any $a \in \mathbb{R}$

$$F_X(a) = \int_{-\infty}^a f_X(y) dy.$$

In this case, we call f_X the probability density function (PDF) of X .

Remark 4.16. Note that

- (1) From Theorem 4.2 we know that F_X determines $\mathbf{P}$, which means that we can use f_X to calculate the probability of any event. For example let $A \in \mathfrak{B}$, then

$$\mathbf{P}_X(A) = \int_A f_X(y) dy,$$

but, although this looks like an easy proposition (indeed it is if A is an open interval), proving this equality for a general $A \in \mathfrak{B}$ is beyond the scope of course.

- (2) We always know that

$$1 = \mathbf{P}_X(\mathbb{R}) = \int_{-\infty}^{\infty} f_X(y) dy.$$

Exercise 4.2. Let X be an absolutely continuous random variable with a probability density function f defined as such

$$f_X(y) = \begin{cases} C(2y - y^2), & y \in [0, 2] \\ 0, & \text{otherwise} \end{cases}.$$

What is the value of C ? What is the probability that $X > 1$?

Solution. In order calculate C we can use (2) from Remark 4.16 and get

$$1 = \int_{-\infty}^{\infty} f_X(y) dy = C \int_0^2 2y - y^2 dy = C \left[y^2 - \frac{y^3}{3} \right]_0^2 = \frac{4}{3}C,$$

which implies that $C = \frac{3}{4}$. In order to find the probability that $X > 1$ we can use (1) from Remark 4.16 and see that

$$\mathbf{P}_X((1, \infty)) = \mathbf{P}_X((1, 2)) = \frac{3}{4} \int_1^2 2y - y^2 dy = \frac{1}{2}.$$

4.8 Uniform distribution

Definition 4.12 (Uniform distribution). Let $[a, b] \subset \mathbb{R}$ be a closed interval. We say that a random variable X is uniformly distributed on $[a, b]$ and write $X \sim U([a, b])$ if it is absolutely continuous with a probability density function

$$f_X(y) = \begin{cases} \frac{1}{b-a}, & y \in [a, b] \\ 0, & \text{otherwise} \end{cases}.$$

Alternatively, X is uniformly distributed on $[a, b]$ if

$$F_X(t) = \begin{cases} 0, & t \leq a \\ \frac{t-a}{b-a}, & a < t \leq b \\ 1, & b < t \end{cases}.$$

Exercise 4.3. A bus reaches a train station every 15 minutes starting at 5:00AM. A student reaches the station in uniform distribution somewhere between 7:00AM and 7:30AM. We describe the time the student reached the station with a random variable X . We have that $X \sim U([a, b])$. What is the probability that the student had to wait more than 5 minutes for the bus?

Solution. Denote by X then time the student reached the bus station. Since we only care about times between 7:00AM and 7:30AM, we can let $X: \Omega \rightarrow [0, 30]$. Since the bus only arrives at 7:00AM, 7:15AM, and 7:30AM, the student will wait more than 5 minutes if and only if $\{X \in (0, 10]\}$ or $\{X \in (15, 25]\}$. Thus,

$$\mathbf{P}_X((0, 10] \cup (15, 25]) = \mathbf{P}_X((0, 10]) + \mathbf{P}_X((15, 25]) = \int_0^{10} \frac{1}{30} dy + \int_{15}^{25} \frac{1}{30} dy = \frac{2}{3}.$$

Definition 4.13 (Exponential distribution). We say that a random variable X is exponentially distributed with parameter $\lambda > 0$, and denote $X \sim \text{Exp}(\lambda)$ if it is absolutely continuous with a probability density function

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

Alternatively, X is exponentially distributed with parameter λ if

$$F_X(t) = \int_0^t \lambda e^{-\lambda y} dy = 1 - e^{-\lambda t}$$

Remark 4.17. The exponential distribution is the continuous version of the geometric distribution we saw in discrete random variables. An exponential random variable is generally a good model of describing the next time an event will happen if the following conditions are satisfied:

- (1) The events occur independently.
- (2) The rate of occurrence is constant.
- (3) Two occurrences can't happen at the same time.

Definition 4.14 (Memorylessness). A continuous random variable X is called memoryless if it satisfies the equality

$$\mathbf{P}(X > a + b \mid X > a) = \mathbf{P}(X > b),$$

for all $a, b \geq 0$. If X is discrete then the definition it is called memoryless if

$$\mathbf{P}(X > a + b \mid X \geq a) = \mathbf{P}(X > b),$$

for all $a, b \geq 0$. Intuitively, if X is the lifespan of someone, This means that the probability they live b additional years does not depend on how old they are right now, and therefore is simply equal to the probability they reach age b (starting from zero).

Remark 4.18. The only discrete memoryless distribution is the geometric distribution, and the only continuous memoryless distribution is the exponential distribution.

4.9 Normal distribution

Definition 4.15 (Normal distribution). We say that a random variable X is normally distributed with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, and write $X \sim N(\mu, \sigma^2)$ if it is absolutely continuous with a probability density function

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

Alternatively, X is normally distributed with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ if

$$F_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^t \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy.$$

In the case of $\mu = 0$ and $\sigma^2 = 1$ we say that it is the *standard normal distribution*. Normal distribution is sometimes called Gaussian distribution.

Remark 4.19. The cumulative distribution function of normal distributions is so hard to calculate in cases where $t \notin \{-\infty, \infty\}$ that there is a special way to denote it. In the case of standard normal distribution:

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

Theorem 4.7. Let $X_{\mu, \sigma^2} \sim N(\mu, \sigma^2)$. Then for all $-\infty \leq a < b \leq \infty$, the probability of:

$$\mathbf{P}(\mu + a\sigma \leq X_{\mu, \sigma^2} \leq \mu + b\sigma)$$

does not depend on μ or σ . In particular, if $Z \sim N(0, 1)$ then:

$$\mathbf{P}(\mu + a\sigma \leq X_{\mu, \sigma^2} \leq \mu + b\sigma) = \mathbf{P}(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Proof. Using the change of variables $z = \frac{y-\mu}{\sigma}$ and $dy = \sigma dz$ we get:

$$\begin{aligned} \mathbf{P}(\mu + a\sigma \leq X_{\mu, \sigma^2} \leq \mu + b\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu+a\sigma}^{\mu+b\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy \\ &= \frac{1}{\sqrt{2\pi}} \int_a^b \exp(-z^2/2) dz = \mathbf{P}(a \leq Z \leq b). \end{aligned}$$

□

Remark 4.20. This is not the last time we are going to hear about the normal distribution (in fact far from it), so even though it is less relevant now, don't forget Theorem 4.7.

4.10 Mixed random variables

Theorem 4.8. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. Then there exists a distribution function of a discrete random variable F_X^d , a distribution function of a continuous random variable F_X^c , and a number $0 \leq \alpha \leq 1$ such that

$$F_X = \alpha F_X^d + (1 - \alpha) F_X^c.$$

Definition 4.16 (Mixed random variable). A random variable X that is not discrete or continuous, is called a mixed random variable.

Example 4.5. Suppose we flip a coin, and when it lands on heads we uniformly choose a number in $[0, 1]$ and write it on a piece of paper, and when it lands on tails we roll a fair die and write the result on a piece of paper. If we let $X: \Omega \rightarrow \mathbb{R}$ be the random variable that sends a result to the number written on the paper, then F_X is not continuous, but it is also not a step function. In this case, X is a mixed random variable, and it can be shown that $F_X = \frac{1}{2}F_X^d + \frac{1}{2}F_X^c$, where F_X^d and F_X^c are the distribution functions of the discrete and continuous experiments respectively.

Remark 4.21. In the scope of this course, each time we decompose a mixed random variable to its discrete and continuous part, we will always assume that the continuous part is absolutely continuous (recall that we don't have the tools to analyze general continuous random variables yet).

4.11 Measurable spaces and measurable functions

Example 4.6. Let $X \sim U([0, 1])$, and denote $Y = X^2$. Is Y a random variable? What is its cumulative distribution function? Is it also uniform?

First we notice that it is a random variable, since for any open interval $(a, b) \subset \mathbb{R}$ we have $Y^{-1}((a, b)) \in \mathcal{F}$. We also have that

$$F_Y(a) = \mathbf{P}(Y \leq a) = \mathbf{P}(X^2 \leq a) = \mathbf{P}(X \leq \sqrt{a}) = F_X(\sqrt{a}).$$

Thus, since $X \sim U([0, 1])$ we get

$$F_Y(x) = \begin{cases} 0, & x < 0 \\ \sqrt{x}, & x \in [0, 1] \\ 1, & x > 1 \end{cases}.$$

This shows that Y does not distribute uniformly.

Remark 4.22 (Generalizing Example 4.6). Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and let $h: \mathbb{R} \rightarrow \mathbb{R}$ be a function. In Example 4.6 we saw that for $h(x) = x^2$, we get that $Y(\omega) = h(X)(\omega) = h(X(\omega)) = X^2(\omega)$ is a random variable with $F_Y(a) = F_X(h^{-1}(a))$. We now want to generalize and find for which h is $h(X)$ a random variable and we want to find a way to describe F_Y using F_X and h .

In order for $Y = h(X)$ to be a random variable we need to require

$$Y^{-1}((a, b)) = \{\omega \in \Omega \mid h(X(\omega)) \in (a, b)\} = \{\omega \in \Omega \mid X(\omega) \in h^{-1}((a, b))\} \in \mathcal{F}.$$

From the recitations we know that if $Y^{-1}((a, b)) \in \mathcal{F}$ for all open intervals $(a, b) \subset \mathbb{R}$, then $Y^{-1}(A) \in \mathcal{F}$ for all $A \in \mathfrak{B}(\mathbb{R})$. This gives rise to the definition of a measurable function.

Definition 4.17 (Measurable space). Consider a set X and a σ -algebra $\mathcal{F}$ on X . Then the tuple $(X, \mathcal{F})$ is called a measurable space.

Definition 4.18 (Measurable function). Let (X, Σ) and (Y, T) be measurable spaces. A function $f: X \rightarrow Y$ is said to be T -measurable if for every $E \in T$ the preimage of E under f is in Σ ; that is, for all $E \in T$

$$f^{-1}(E) := \{x \in X \mid f(x) \in E\} \in \Sigma.$$

Example 4.7 (Borel measurable function). A function $f: (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathfrak{B})$ so that $f^{-1}(A) \in \mathcal{F}$ for all $A \in \mathfrak{B}(\mathbb{R})$ is called a $\mathfrak{B}$ -measurable, or a Borel measurable function. As mentioned earlier, it suffices to check measurability of preimages of open intervals.

Definition 4.19 (σ -algebra generated by a function). Let $f: (X, \Sigma) \rightarrow (Y, T)$ be a function between measure spaces. Then the σ -algebra generated by f , denoted $\sigma(f) \subseteq \Sigma$, is given by

$$\sigma(f) := \{f^{-1}(E) \mid E \in T\}.$$

Remark 4.23. The σ -algebra $\sigma(f)$ is the smallest σ -algebra that makes f measurable.

Proposition 4.9. Let $X: \Omega \rightarrow \mathbb{R}$ be a random variable, and let $h: \mathbb{R} \rightarrow \mathbb{R}$ be a Borel measurable function, then $Y = h(X)$ is a random variable.

Theorem 4.10. Let $h: \mathbb{R} \rightarrow \mathbb{R}$ be a function with a countable amount of discontinuities $\{a_i\}_{i \in \mathbb{N}}$. If $\{a_i\}_{i \in \mathbb{N}}$ has no accumulation point in $\mathbb{R}$, then h is Borel measurable.

Proposition 4.11 (Distribution function of $Y = h(X)$). Let X be a random variable, and let h be a measurable function. Then

- (1) If h is a strictly increasing function, then $F_Y(a) = F_X(h^{-1}(a))$, where if h is not surjective (invertible) we define

$$h^{-1}(a) := \sup \{b \in \mathbb{R} \mid h(b) \leq a\}.$$

Additionally, if X is an absolutely continuous variable and h is differentiable, then Y is absolutely continuous and $f_Y(a) = f_X(h^{-1}(a)) \cdot |(h^{-1})'(a)|$.

- (2) If h is a strictly decreasing function then $F_Y(a) = 1 - \lim_{y \rightarrow h^{-1}(a)^-} F_X(y)$, where if h is not surjective (invertible) we define

$$h^{-1}(a) := \inf \{b \in \mathbb{R} \mid h(b) \geq a\}.$$

Additionally, if X is an absolutely continuous variable and h is differentiable, then Y is absolutely continuous, and $f_Y(a) = f_X(h^{-1}(a)) \cdot |(h^{-1})'(a)|$.

Proof. Immediate from the definitions and similar to what we did in Example 4.6. $\square$

Remark 4.24. Theorem 4.11 may look a bit intimidating since without the intuition things can look very complicated. I cannot explain the intuition easily here, but I will say that this is exactly *pushforward*, the mathematical twin of the *pullback*, which is in turn a generalization of the change of variables theorem (the condition that it is strictly increasing is the condition that the function is invertible). The concept of change of variables is taught comprehensively in calculus 3 (usually not very well), and on special occasion you may also get to learn about *pullback* there. These are definitely not simple concepts but understanding them well is very useful everywhere. For readers who have read a bit about measure theory and have not taken a formal course, note that the density functions are not the measures themselves.

Remark 4.25. Theorem 4.11 gave us a way to calculate the distribution of $Y = h(X)$ when h is strictly monotonic. We would also like to compute F_Y when h is a more complex function like a polynomial. In that case we partition the domain of h into intervals in which it is monotonic.

Exercise 4.4. Let $Y: \Omega \rightarrow \mathbb{R}$ be the distance of a point uniformly chosen on the upper half of the unit circle from the x axis. Compute F_Y .

Solution. Let $X(\omega)$ be the angle that the point creates with the x axis. Then the point is $(\cos X, \sin X)$, and in particular $Y = \sin X$. The function $t \mapsto \sin t$ is not monotonic on $[0, \pi]$, however it is monotonic on $[0, \pi/2]$ and $[\pi/2, \pi]$ separately. Therefore we get

$$\begin{aligned} F_Y(a) &= \mathbf{P}(Y \leq a) = \mathbf{P}(\sin X \leq a) \\ &= \mathbf{P}(\sin X \leq a \text{ and } X \in [0, \pi/2]) + \mathbf{P}(\sin X \leq a \text{ and } X \in (\pi/2, \pi]) \\ &= \mathbf{P}(X \in [0, \arcsin a]) + \mathbf{P}(X \in (\pi - \arcsin a, \pi]) \\ &= \frac{\arcsin a}{\pi} + \mathbf{P}(X \in (\pi - \arcsin a, \pi]). \end{aligned}$$

In order to compute $\mathbf{P}(\sin X \leq a \text{ and } X \in (\pi/2, \pi])$ we can either use symmetry or compute directly

$$\mathbf{P}(\sin X \leq a \text{ and } X \in (\pi/2, \pi]) = F_X(\pi) - F_X(\pi - \arcsin a) = 1 - \frac{\pi - \arcsin a}{\pi} = \frac{\arcsin a}{\pi}.$$

Thus,

$$F_Y(a) = \frac{2 \arcsin a}{\pi}.$$

4.12 Simulation of random variables

Proposition 4.12. *Let X be an absolutely continuous random variable with strictly increasing distribution function $F_X(x)$ on $[a, b]$ such that $F_X(a) = 0$ and $F_X(b) = 1$ where $-\infty \leq a < b \leq \infty$. Define $U = F_X(X)$. Then $U \sim U([0, 1])$.*

Proof. Since $F_X(a) = 0$ and $F_X(b) = 1$ we must have that:

$$F_U(y) = \begin{cases} 0, & y \leq 0 \\ 1, & y \geq 1 \end{cases}$$

and for $0 \leq y \leq 1$ we have that:

$$F_U(y) = \mathbf{P}(U \leq y) = \mathbf{P}(F_X(X) \leq y) = \mathbf{P}(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y,$$

and thus the proof is complete. We may also note that $f_U(y) = F'_U(y) = 1$ iff $0 \leq y \leq 1$ and $f_U(y) = 0$ otherwise. $\square$

Proposition 4.13. *Let F be a strictly increasing continuous distribution function on $(-\infty, \infty)$. Let $U \sim U([0, 1])$ and let $X = F^{-1}(U)$. Then $F_X = F$.*

Proof. We can immediately see that:

$$F_X(x) = \mathbf{P}(X \leq x) = \mathbf{P}(F^{-1}(U) \leq x) = \mathbf{P}(U \leq F(x)) = F_U(F(x)) = F(x).$$

$\square$

Example 4.8 (Simulating a variable). In order to simulate a simple absolutely continuous random variable $U \sim U([a, b])$, we can define the sequence $U_n \sim \text{Bin}(n, p)$. In this way, by considering each result of n trials as a binary number $(0.a_1a_2 \dots a_n)_2$ where a_i is 1 if and only if the i th trial was a success and zero otherwise, we clearly get that U_n converges to U both in distribution and almost surely (we will learn what these mean in Section 12.1).

In case we have a more general random variable X with a strictly increasing F_X then we can use Theorem 4.13 and have that $X = F_X^{-1}(U)$ and we get that the variables $V_n = F_X^{-1}(U_n)$ converge to X .

5 Expectation, variance and moments

5.1 Expectation

Remark 5.1. The expectation of a random variable represents the weighted “average” of its values.

Definition 5.1 (Expectation, discrete case). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X: \Omega \rightarrow \mathbb{R}$ be a discrete random variable. Suppose that $\text{supp}(\mathbf{P}_X) = \{a_i\}_{i=1}^\infty$, then we define

$$\begin{aligned}\mathbf{E}_+[X] &= \sum_{i: a_i \geq 0} a_i \mathbf{P}_X(\{a_i\}) \\ \mathbf{E}_-[X] &= \sum_{i: a_i \leq 0} a_i \mathbf{P}_X(\{a_i\}).\end{aligned}$$

If at least one of $\mathbf{E}_+[X]$, $\mathbf{E}_-[X]$ are finite, we define

$$\mathbf{E}[X] = \mathbf{E}_+[X] + \mathbf{E}_-[X].$$

If either one of $\mathbf{E}_+[X]$, $\mathbf{E}_-[X]$ are infinite, we say that $\mathbf{E}[X]$ is undefined.

Definition 5.2 (Expectation, absolutely continuous case). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, If X is an absolutely continuous variable with a probability density function f_X , then we define

$$\begin{aligned}\mathbf{E}_+[X] &= \int_0^\infty y f_X(y) dy \\ \mathbf{E}_-[X] &= \int_{-\infty}^0 y f_X(y) dy\end{aligned}$$

If at least one of $\mathbf{E}_+[X]$, $\mathbf{E}_-[X]$ are finite, we define

$$\mathbf{E}[X] = \mathbf{E}_+[X] + \mathbf{E}_-[X].$$

If either one of $\mathbf{E}_+[X]$, $\mathbf{E}_-[X]$ are infinite, we say that $\mathbf{E}[X]$ is undefined.

Remark 5.2. If X is a mixed random variable, such that $F_X = \alpha F_X^d + (1 - \alpha) F_X^{ac}$ then

$$\mathbf{E}[X] = \alpha(\mathbf{E}_+[X_d] + \mathbf{E}_-[X_d]) + (1 - \alpha)(\mathbf{E}_+[X_{ac}] + \mathbf{E}_-[X_{ac}]),$$

assuming either $\mathbf{E}_+[X_d], \mathbf{E}_+[X_{ac}] < \infty$ or $\mathbf{E}_-[X_d], \mathbf{E}_-[X_{ac}] < \infty$.

Example 5.1. Let $X \sim \text{Bin}(n, p)$. Then

$$\begin{aligned}\mathbf{E}[X] &= \sum_{k=0}^n k \mathbf{P}_X(k) \\ &= \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^n \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^n \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} \\ &= np(p + (1-p))^{n-1} \\ &= np.\end{aligned}$$

Remark 5.3. In order to solve questions it is necessary to know the expectation of any random variable which distributes in the standard ways we discussed before (for example for $X \sim N(\mu, \sigma^2)$ etc.). I will not cover all the calculations here, but the results of such expectations are always given in the formula sheet for the test.

Proposition 5.1. Let $X: \Omega \rightarrow [0, \infty)$ be a nonnegative random variable. Then

$$\mathbf{E}[X] = \int_0^\infty 1 - F_X(t) dt.$$

Moreover, if $X: \Omega \rightarrow \mathbb{R}$ is a general random variable with a well defined expectation then

$$\mathbf{E}[X] = \int_0^\infty 1 - F_X(t) - F_X(-t) dt.$$

Remark 5.4 (Survival function intuition). In Theorem 5.1, we saw the expression $S_X(t) = 1 - F_X(t)$, which is usually considered the survival function. That is because if $X(t)$ is our probability to live to age t , then $S_X(t) = 1 - F_X(t) = 1 - F_X(X \leq t) = F_X(X > t)$ is the probability we will survive after age t . Moreover, in the case X is nonnegative, it is very intuitive that the average life expectation $\mathbf{E}[X]$, is equal to integrating the probability of living after a certain age $\int_0^\infty S_X(t) dt$. Similarly, in the general case, we expect that $\mathbf{E}[X] = \int_{-\infty}^\infty S_X(t) dt$, or equivalently $\mathbf{E}[X] = \int_0^\infty S_X(t) - S_{-X}(t) dt$, and then since $S_{-X}(t) = \mathbf{P}_X(-X > t) = \mathbf{P}_X(X < -t) = F_X(-t)$, we get $\mathbf{E}[X] = \int_0^\infty 1 - F_X(t) - F_X(-t) dt$.

Remark 5.5. Given a random variable X and a measurable function $h: \mathbb{R} \rightarrow \mathbb{R}$, we saw that we can describe the probability distribution function $F_{h(X)}$ by F_X , and we now saw that we can describe $\mathbf{E}[X]$ by F_X , so it makes sense to ask if there is a relation between $\mathbf{E}[h(X)]$ and the distribution of X . The following two propositions are going to show that relation; they are known as the “law of the unconscious statistician” or LOTUS for short.

Example 5.2. Let $X \sim \text{Exp}(\lambda)$. Since X is nonnegative (also denoted $X \geq 0$), and for $t > 0$ we have $F_X(t) = 1 - e^{-\lambda t}$, we get

$$\mathbf{E}[X] = \int_0^\infty 1 - F_X(t) dt = \int_0^\infty e^{-\lambda t} dt = \frac{1}{\lambda}.$$

Proposition 5.2 (LOTUS, discrete case). Let X be a discrete random variable with support denoted $\text{supp}(\mathbf{P}_X) = \{a_i\}_{i=1}^\infty$, and let $h: \mathbb{R} \rightarrow \mathbb{R}$ be a measurable function. If $\sum_{i=1}^\infty h(a_i) \mathbf{P}_X(a_i) < \infty$ then we have:

$$\mathbf{E}[h(X)] = \sum_{i=1}^\infty h(a_i) \mathbf{P}_X(a_i).$$

Proposition 5.3 (LOTUS, absolutely continuous case). Let X be an absolutely continuous random variable with a probability density function f_X , and let $h: \mathbb{R} \rightarrow \mathbb{R}$ be Borel measurable. If $\int_{-\infty}^\infty |h(x)| f_X(x) dx < \infty$ then we have:

$$\mathbf{E}[h(X)] = \int_{-\infty}^\infty h(x) f_X(x) dx.$$

Remark 5.6. These propositions are called the “law of the unconscious statistician”, because people supposedly assume these are definitions, instead of propositions that are based on the true definition of expectation. Indeed, these are the nicest formulas we could have hoped for.

Exercise 5.1. In a certain island, the population increases by 5% each year. In the first year, the population of the island is 1000 people, but each year a coin is flipped (with probability p to land on heads), and if it lands on tails nothing happens. If it lands on heads, all the people on the island die a painful death. What is the expectation of the amount of people who are going to die?

Solution. We can notice that the number of years the population will survive distributes geometrically so we denote $N \sim \text{Geo}(p)$. We know that the number of people on the island after t years (without the coin landing on heads) is

$$h(t) = 1000 \cdot \left(1 + \frac{5}{100}\right)^t.$$

Let $X = h(N)$ denote the number of people on the island. Using the LOTUS formula for a discrete variable we get

$$\begin{aligned} \mathbf{E}[X] &= 1000 \sum_{n=1}^{\infty} \left(1 + \frac{5}{100}\right)^n \mathbf{P}_N(n) = 1000 \sum_{n=1}^{\infty} \left(\frac{105}{100}\right)^n (1-p)^{n-1} p \\ &= 1000 \cdot \frac{105}{100} \cdot p \sum_{n=1}^{\infty} \left(\frac{105}{100}(1-p)\right)^{n-1}. \end{aligned}$$

Since the final series is a geometric series we conclude that

$$\mathbf{E}[X] = \begin{cases} 1050p \cdot \frac{1}{1 - \frac{105}{100}(1-p)}, & \frac{105}{100}(1-p) < 1 \\ \infty, & \frac{105}{100}(1-p) \geq 1 \end{cases}$$

Exercise 5.2. What is the expected area of a square with side length $X \sim U([0, 1])$?

Solution. By the LOTUS formula

$$\mathbf{E}[X^2] = \int_{-\infty}^{\infty} y^2 f_X(y) dy = \int_0^1 y^2 dy = \frac{1}{3}.$$

Proposition 5.4 (Properties of expectation). *Let X be a random variable, let $a \in \mathbb{R}$, and let $h_1, h_2: \mathbb{R} \rightarrow \mathbb{R}$ be Borel measurable functions. Then*

- (1) $\mathbf{E}[aX] = a\mathbf{E}[X]$.
- (2) $\mathbf{E}[h_1(X) + h_2(X)] = \mathbf{E}[h_1(X)] + \mathbf{E}[h_2(X)]$.
- (3) If $\mathbf{P}(X = a) = 1$, then $\mathbf{E}[X] = a$.
- (4) If $\mathbf{P}(h_1(X) \geq 0) = 1$, then $\mathbf{E}[h_1(X)] \geq 0$.

Assuming all expectations are well defined and finite, from (2) and (3) we get that $\mathbf{E}[h_1(X) + a] = \mathbf{E}[h_1(X)] + a$.

Proof. The proofs here are still done for each type of variable separately, and they aren't too complicated so I choose to omit them. $\square$

5.2 Variance and standard deviation

Definition 5.3 (Variance and standard deviation). Let X be a random variable with finite expectation. The Variance of X , denoted $\text{Var}(X)$, is defined as

$$\text{Var}(X) := \mathbf{E} \left[(X - \mathbf{E}[X])^2 \right].$$

The standard deviation of X , denoted σ_X , is defined as

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

Remark 5.7. Intuitively, we say that $\text{Var}(X)$ and σ_X describe how far is X from its expected value on average. Another expression that describes this is $\mathbf{E}[|X - \mathbf{E}[X]|]$ but we don't use it for a couple of reasons. First, in many cases it's easier to calculate the variance and standard deviation. Secondly, the standard deviation σ_X is the probabilistic version of Euclidean distance, and the variance $\text{Var}(X)$ describes that square of the Euclidean distance, which are both very natural parameters, while $\mathbf{E}[|X - \mathbf{E}[X]|]$ isn't as natural.

Example 5.3. If we denote the result of a dice throw by X , then we clearly see that $\mathbf{E}[X] = 3.5$, so the variance is

$$\text{Var}(X) = \mathbf{E}\left[(X - 3.5)^2\right] = \sum_{i=1}^6 (i - 3.5)^2 \cdot \frac{1}{6} = \frac{35}{12},$$

where the middle transition is an application of the law of the unconscious statistician.

Remark 5.8. We have that

$X \sim$	$\text{Bin}(n, p)$	$\text{Geo}(p)$	$\text{Pois}(\lambda)$	$U([a, b])$	$\text{Exp}(\lambda)$	$N(\mu, \sigma^2)$
$\text{Var}(X)$	$np(1-p)$	$\frac{1-p}{p^2}$	λ	$\frac{(b-a)^2}{12}$	$\frac{1}{\lambda^2}$	σ^2

Proposition 5.5 (Properties of variance). *Let X be a random variable with finite expectation, and let $a \in \mathbb{R}$. Then*

- (1) $\text{Var}(X) \geq 0$.
- (2) $\text{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$.
- (3) $\text{Var}(X + a) = \text{Var}(X)$.
- (4) $\text{Var}(aX) = a^2 \text{Var}(X)$.
- (5) $\text{Var}(X) = 0$ if and only if $\mathbf{P}(X = \mathbf{E}[X]) = 1$.

In particular from (1) and (2) we get that $\mathbf{E}[X]^2 \leq \mathbf{E}[X^2]$.

Proof. The proofs are immediate from properties of expectation.

- (1) By definition $\text{Var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2]$, and we know that $(X - \mathbf{E}[X])^2 \geq 0$, so from properties of the expectation $\text{Var}(X) \geq 0$.
- (2) By using properties of expectation we get

$$\begin{aligned} \text{Var}(X) &= \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2 - 2\mathbf{E}[X] \cdot X + \mathbf{E}[X]^2] \\ &= \mathbf{E}[X^2] + \mathbf{E}[-2\mathbf{E}[X] \cdot X] + \mathbf{E}[\mathbf{E}[X]^2] \\ &= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}[X]^2 \\ &= \mathbf{E}[X^2] - \mathbf{E}[X]^2. \end{aligned}$$

- (3) By linearity of expectation

$$\begin{aligned} \text{Var}(X + a) &= \mathbf{E}[(X + a - \mathbf{E}[X + a])^2] = \mathbf{E}[(X + a - \mathbf{E}[X])^2 - a^2] \\ &= \mathbf{E}[(X - \mathbf{E}[X])^2] = \text{Var}(X). \end{aligned}$$

- (4) Similar to (4).
- (5) This property takes some effort, we need to show it for discrete and absolutely continuous variables separately.

□

5.3 Moments

Definition 5.4 (Moments). Let X be a random variable.

- (1) The n th moment of X for $n \in \mathbb{N}$ is denoted $\mu_n(X)$ or $m_n(X)$ and is defined by

$$\mu_n(X) := \mathbf{E}[X^n],$$

when the expectation is well defined. In particular $\mu_1(X) = m_1(X) = \mathbf{E}[X]$.

- (2) If X is a nonnegative random variable, then the $\alpha \in \mathbb{R}$ moment of X is defined by

$$\mu_\alpha(X) := \mathbf{E}[X^\alpha].$$

- (3) The n th central moment of X for $n \in \mathbb{N}$ is denoted $C_n(X)$ and is defined as

$$C_n(X) := \mathbf{E}[(X - \mathbf{E}[X])^n].$$

In particular, $C_0(X) = 1$, $C_1(X) = \mathbf{E}[X] - \mathbf{E}[X] = 0$, and $C_2(X) = \text{Var}(X)$.

Example 5.4. Let $X \sim U([a, b])$, then

$$\mu_n(X) = m_n(X) = \mathbf{E}[X^n] = \int_a^b \frac{y^n}{b-a} dy = \frac{b^{n+1} - a^{n+1}}{(n+1)(b-a)}.$$

Definition 5.5 (Moment generating function). Let X be a random variable. The moment generating function of X , denoted $M_X: \mathbb{R} \rightarrow [0, \infty]$ is defined by

$$M_X(t) := \mathbf{E}[e^{tX}].$$

Remark 5.9. Notice that M_X is always defined since $y \mapsto e^{ty}$ is a positive function for all real values $t \in \mathbb{R}$.

Remark 5.10. The relation between the moment generating function M_X and the moments becomes clearer when we recall that the series expansion of e^{tX} is $\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}$, and when we take $M_X(t) = \mathbf{E}[e^{tX}]$ we get the series $M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbf{E}[X^n] = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mu_n(X)$. This fact immediately implies Theorem 5.6. However, before trying to understand the theory of moments too deeply, I suggest first treating them as computational tools.

Example 5.5. Let $X \sim \text{Bin}(n, p)$, then

$$M_X(t) = \mathbf{E}[e^{tX}] = \sum_{k=0}^n e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = (pe^t + 1 - p)^n.$$

Proposition 5.6. Let X be a random variable, and suppose that M_X is finite in $[-\epsilon, \epsilon]$ for some $\epsilon > 0$. Then M_X is a smooth function and for all $k \geq 0$ we get

$$M_X^{(k)}(0) = \mu_k(X) = \mathbf{E}[X^k].$$

Remark 5.11. There are cases where it is easier to calculate the moment generating function than to calculate the moments themselves. Using the previous proposition, under the conditions of the proposition, we can calculate the moments by differentiating M_X .

Remark 5.12. Moments have many uses, for example, we can use moments to find bounds for events of the form $\{X > a\}$ or $\{|X - \mathbf{E}[X]| > a\}$.

Proposition 5.7. Let X and Y be random variables. If there exists $\epsilon > 0$ such that M_X and M_Y are finite on $[0, \epsilon]$ and $M_X(t) = M_Y(t)$ for $t \in [0, \epsilon]$, then $\mathbf{P}_X = \mathbf{P}_Y$.

Remark 5.13. The intuitive meaning of Theorem 5.7, is that if X and Y have identical finite moments, and the functions $n \mapsto \mathbf{E}[X^n]$, $n \mapsto \mathbf{E}[Y^n]$ do not grow “too fast”, then they have the same probability distribution $\mathbf{P}_X = \mathbf{P}_Y$.

6 Important inequalities

6.1 Jensen's inequality

Remark 6.1. Before proving Jensen's inequality we will start by proving a less general result which is also known as Jensen's inequality.

Proposition 6.1. *Let X be a random variable, then*

$$|\mathbf{E}[X]| \leq \mathbf{E}[|X|].$$

Proof. In the case of a discrete random variable with $\text{supp}(\mathbf{P}_X) = \{a_i\}_{i \in \mathbb{N}}$ we get

$$|\mathbf{E}[X]| = \left| \sum_{i=1}^{\infty} a_i P_X(a_i) \right| \leq \sum_{i=1}^{\infty} |a_i| P_X(a_i) = \sum_{i=1}^{\infty} |a_i| P_X(a_i) = \mathbf{E}[|X|].$$

Similarly, if X is an absolutely continuous random variable with a probability density function f_X then

$$|\mathbf{E}[X]| = \left| \int_{-\infty}^{\infty} y f_X(y) dy \right| \leq \int_{-\infty}^{\infty} |y| f_X(y) dy = \int_{-\infty}^{\infty} |y| f_X(y) dy = \mathbf{E}[|X|].$$

The proof for a mixed random variable is very similar. □

Remark 6.2. Previously we also saw that $\mathbf{E}[X]^2 \leq \mathbf{E}[X^2]$. Both this inequality and Theorem 6.1 are inequalities of the form

$$h(\mathbf{E}[X]) \leq \mathbf{E}[h(X)].$$

In the first inequality h is the function $h(y) = y^2$ and in the second it is $h(y) = |y|$. We want to find for which functions the inequality is satisfied, and for this purpose we have the general Jensen's inequality.

Theorem 6.2 (Jensen's inequality). *Let X be a random variable with finite expectation, and let $h: \mathbb{R} \rightarrow \mathbb{R}$ a convex function. Then*

$$h(\mathbf{E}[X]) \leq \mathbf{E}[h(X)].$$

Remark 6.3. Recall that a function $h: \mathbb{R} \rightarrow \mathbb{R}$ is said to be convex if for every $x, y \in \mathbb{R}$ and any $0 \leq t \leq 1$ we have

$$h(tx + (1-t)y) \leq th(x) + (1-t)h(y).$$

Remark 6.4. Additionally, recall from analysis 1, that if $h: \mathbb{R} \rightarrow \mathbb{R}$ is convex, then for any $x_0 \in \mathbb{R}$ there exists a linear function $f_{x_0}(y) = ay + b$, or equivalently $f_{x_0}(y) = a(y - x_0) + h(x_0)$, so that $f_{x_0}(y) \leq h(y)$ for all $y \in \mathbb{R}$.

Proof of Theorem 6.2. Denote $x_0 := \mathbf{E}[X]$, and let $f_{x_0}(y) = a(y - x_0) + h(x_0)$ be the corresponding linear function as in Remark 6.4. For all $y \in \mathbb{R}$ we get

$$h(y) \geq f_{x_0}(y) = a(y - x_0) + h(x_0) = a(y - \mathbf{E}[X]) + h(\mathbf{E}[X]).$$

Since this holds for all $y \in \mathbb{R}$ we conclude that

$$h(X) \geq a(X - \mathbf{E}[X]) + h(\mathbf{E}[X]).$$

Now from properties of expectation we get

$$\begin{aligned} \mathbf{E}[h(X)] &\geq \mathbf{E}[a(X - \mathbf{E}[X]) + h(\mathbf{E}[X])] = \mathbf{E}[aX] - \mathbf{E}[a\mathbf{E}[X]] + \mathbf{E}[h(\mathbf{E}[X])] \\ &= a\mathbf{E}[X] - a\mathbf{E}[X] + h(\mathbf{E}[X]) = h(\mathbf{E}[X]), \end{aligned}$$

which completes the proof. □

Remark 6.5. If $I \subset \mathbb{R}$ is an interval so that $\mathbf{P}(X \in I) = 1$, and $h: I \rightarrow \mathbb{R}$ is convex then as we may expect Jensen's inequality also gives $h(\mathbf{E}[X]) \leq \mathbf{E}[h(X)]$.

Remark 6.6. If $h: \mathbb{R} \rightarrow \mathbb{R}$ is a concave function, then $-h: \mathbb{R} \rightarrow \mathbb{R}$ is a convex function, and therefore by Jensen's inequality and linearity of expectation $h(\mathbf{E}[X]) \geq \mathbf{E}[h(X)]$.

Remark 6.7 (Intuition for Jensen's inequality). The intuition behind Jensen's inequality can be seen by a simple example. If you ever forget Jensen's inequality, you can just recall this example. Suppose that a gamemaster shows you two games. In the first game, you roll a fair die and get result X from 1 to 6, then your prize amount is $\varphi(X)$ gold coins, where $\varphi(x) = x^2$ (this is a convex function). In the second game you roll a fair die with unique sides which show the numbers $1^2, 2^2$, all the way up to 6^2 . In the first game your expected award is $\mathbf{E}[\varphi(\mathbf{E}[X])] = \varphi(\mathbf{E}[X])$, and in the second game it is $\mathbf{E}[\varphi(X)]$. Then, Jensen's inequality becomes the question, "what game should you play?". It is clear that in the first game we are expected to get $(3.5)^2$ coins, however, since the numbers on the dice of the second game are squared, their growth rate is increasing (this is exactly saying the function is convex), and thus the bigger numbers have more weight (this logic does not work for the first game because there we take the expected value before we take the square). If this explanation was confusing, just notice that the average of 1, 4, 9, 16, 25, 36 is bigger than $(3.5)^2$, so the second game is more profitable, which means that $\varphi(\mathbf{E}[X]) \leq \mathbf{E}[\varphi(X)]$.

Example 6.1. Let $X \sim U([3, 7])$, and $h(y) = \frac{1}{y}$. Using Jensen's inequality we get that

$$\mathbf{E} \left[\frac{1}{X} \right] \geq \frac{1}{\mathbf{E}[X]}.$$

Indeed,

$$\mathbf{E} \left[\frac{1}{X} \right] = \int_3^7 \frac{1}{y} \cdot \frac{1}{4} dy = \frac{1}{4} \log(7/3) \approx 0.212,$$

while

$$\frac{1}{\mathbf{E}[X]} = \frac{1}{5} = 0.2.$$

Corollary 6.3. Let X be a random variable and $0 < \alpha < \beta$. If $\mu_\beta(X)$ is well defined, that is, $\mu_\beta(X) = \mathbf{E}[|X|^\beta] < \infty$, then $\mu_\alpha(X)$ is also well defined.

Proof. Since $0 < \frac{\alpha}{\beta} < 1$ the function $y \mapsto y^{\alpha/\beta}$ is concave on $[0, \infty)$, and thus from Jensen's inequality for concave functions:

$$\mathbf{E}[|X|^\alpha] = \mathbf{E}[(|X|^\beta)^{\alpha/\beta}] \leq \mathbf{E}[|X|^\beta]^{\alpha/\beta} < \infty,$$

which completes the proof. □

6.2 Markov's inequality

Unlike Jensen's inequality, which was an inequality between expectations, this inequality is about bounding the probability of an event with expectation of the variable. Specifically, we bound the probability of the event $X > a$ (when X is nonnegative), using the expectation of X . This is very intuitive, because if there is a high probability of X being a large number, the expectation should increase accordingly.

Theorem 6.4 (Markov's inequality). Let X be a nonnegative random variable. Then for all $a > 0$ we have

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}.$$

Proof. Let $a > 0$. Define the functions $h(y) := y$ and

$$g(y) := a \cdot \mathbb{1}_{(a, \infty)}(y) = \begin{cases} 0, & y < a \\ a, & y \geq a \end{cases}.$$

Since $g(y) \leq h(y)$ for any $y \geq 0$, and X is a nonnegative random variable ($\mathbf{P}(X \geq 0) = 1$), we get from properties of expectation that

$$\mathbf{E}[X] = \mathbf{E}[h(X)] \geq \mathbf{E}[g(X)] = a \cdot \mathbf{P}(X \geq a),$$

which completes the proof. $\square$

Remark 6.8. Algebraically we can see that the inequality is $a \cdot \mathbf{P}(X \geq a) \leq \mathbf{E}[X]$ which is exactly what we would expect if the variable is nonnegative.

Remark 6.9. Since for any random variable X , the random variable $|X|$ is nonnegative, we get the following corollary for free.

Corollary 6.5. *Let X be a random variable. Then for all $a > 0$ we have*

$$\mathbf{P}(|X| \geq a) \leq \frac{\mathbf{E}[|X|]}{a}.$$

Proposition 6.6 (Markov's inequality for functions). *Let X be a nonnegative random variable and let $\varphi: [0, \infty) \rightarrow [0, \infty)$ be a strictly increasing function. Then for all $a > 0$ we have*

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[\varphi(X)]}{\varphi(a)}.$$

In particular, for any random variable X , we have

$$\mathbf{P}(|X| \geq a) \leq \frac{\mathbf{E}[\varphi(|X|)]}{\varphi(a)}.$$

Proof. We are going to apply Markov's inequality for $\varphi(X)$. First, since φ is increasing, φ is Borel measurable, so $\varphi(X)$ is a random variable. Since φ is strictly increasing we also have $\{X \geq a\} = \{\varphi(X) \geq \varphi(a)\}$. Since X is nonnegative $\varphi(X)$ must also be nonnegative. Thus, from Markov's inequality

$$\mathbf{P}(X \geq a) = \mathbf{P}(\varphi(X) \geq \varphi(a)) \leq \frac{\mathbf{E}[\varphi(X)]}{\varphi(a)},$$

which completes the proof. $\square$

Remark 6.10. Note that some quick applications of Markov's inequality show that $\mathbf{P}(|X| \geq a) \leq \frac{\mathbf{E}[|X|^n]}{a^n}$ and $\mathbf{P}(|X| \geq a) \leq \frac{\mathbf{E}[e^{t|X|}]}{e^{ta}}$ for all $a, t > 0$. This means in particular that variables with well defined high moments have tails that converge to 0 very quickly.

6.3 Chebyshev's inequality

Unlike Jensen's inequality, which was an inequality between expectations, this inequality is about bounding the probability of an event with variation. The intuition is similar to Markov's inequality. As the probability of $|X - \mathbf{E}[X]| \geq a$ increases, the variance increases as well. Since variance is $\mathbf{E}[|X - \mathbf{E}[X]|^2]$, it also makes sense to divide by a^2 instead of a this time.

Theorem 6.7 (Chebyshev's inequality). *Let X be a random variable with finite expectation. Then for every $a > 0$ we have that*

$$\mathbf{P}(|X - \mathbf{E}[X]| \geq a) \leq \frac{\text{Var}(X)}{a^2}.$$

Proof. Using Markov's inequality on the nonnegative random variable $|X - \mathbf{E}[X]|$ with the function $\varphi(y) = y^2$ we get

$$\mathbf{P}(|X - \mathbf{E}[X]| \geq a) = \mathbf{P}(|X - \mathbf{E}[X]|^2 \geq a^2) \leq \frac{\mathbf{E}[|X - \mathbf{E}[X]|^2]}{a^2} = \frac{\text{Var}(X)}{a^2}.$$

□

Example 6.2. Let $X \sim U([0, 2])$. Thus $\mathbf{P}(|X - 1| \geq 1) = 0$. From Chebyshev's inequality we get that

$$\mathbf{P}(|X - 1| \geq 1) \leq \frac{\text{Var}(X)}{1^2} = \frac{1}{3}.$$

6.4 The weak law of large numbers

Chebyshev's inequality allows us to prove a special case of the strong law of large numbers.

Theorem 6.8 (Weak law of large numbers, binomial case). *Let $p \in [0, 1]$. For all $n \in \mathbb{N}$ define $X_n \sim \text{Bin}(n, p)$. Then, for any $\delta > 0$ we have*

$$\lim_{n \rightarrow \infty} \mathbf{P}\left(\left|\frac{X_n}{n} - p\right| > \delta\right) = 0.$$

Proof. Considering the table of values for expectation and variance of $X_n \sim \text{Bin}(n, p)$ we can compute that

$$\begin{aligned} \mathbf{E}[n^{-1}X_n] &= n^{-1}\mathbf{E}[X_n] = n^{-1} \cdot np = p, \\ \text{Var}(n^{-1}X_n) &= n^{-2}\text{Var}(X_n) = n^{-2} \cdot np(1-p) = n^{-1}p(1-p). \end{aligned}$$

From Chebyshev's inequality for the random variable $n^{-1}X_n$, we can now conclude that

$$\begin{aligned} \limsup_{n \rightarrow \infty} \mathbf{P}\left(\left|\frac{X_n}{n} - p\right| > \delta\right) &= \limsup_{n \rightarrow \infty} \mathbf{P}\left(\left|\frac{X_n}{n} - \mathbf{E}[n^{-1}X_n]\right| > \delta\right) \\ &\leq \limsup_{n \rightarrow \infty} \frac{\text{Var}(n^{-1}X_n)}{\delta^2} = \limsup_{n \rightarrow \infty} \frac{p(1-p)}{\delta^2 n} = 0. \end{aligned}$$

□

Remark 6.11. The intuitive meaning of the weak law of large number is that in a repeating sequence of n Bernoulli trials, for any margin $\delta > 0$ we take, as we take $n \rightarrow \infty$, the number of successes divided by n (the number of experiments) converges to p . For example, if the Bernoulli trial is flipping a fair coin, the weak law of large numbers says that the number of times we get heads divided by the number of times we flipped the coin converges to $\frac{1}{2}$ as we flip to coin more and more times.

6.5 The Weierstrass approximation theorem

Theorem 6.9 (Weierstrass approximation theorem). *Let $[a, b] \subset \mathbb{R}$ be a compact interval. Then, for every continuous function $f: [a, b] \rightarrow \mathbb{R}$ there exists a polynomial sequence P_n that converges uniformly to f . In other words*

$$\lim_{n \rightarrow \infty} \sup_{x \in [a, b]} |f(x) - P_n(x)| = 0.$$

Proof. There is a proof based on the weak law of large numbers, but there is also a proof of this theorem in functional analysis, and there is the Stone–Weierstrass’s theorem, which generalizes this result for general function algebras. $\square$

7 Bertrand's paradox

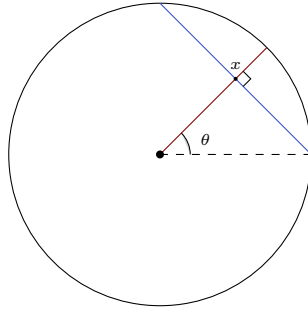
Before continuing to multivariate random variables, let's discuss Bertrand's paradox once again, and solve it. Recall the paradox:

“Consider an equilateral triangle inscribed in a circle. Suppose a chord of the circle is chosen uniformly at random. What is the probability that the chord is longer than a side of the triangle?”

The reason we get a different answer each time is because we are choosing different probability spaces to describe the problem. In the following solutions, for the sake of simplicity we assume that the radius of the circle is 1, and then the length of a side of the triangle is $\sqrt{3}$.

7.1 First solution

In this experiment we describe the chord by choosing an arbitrary angle θ , and a point x on the radius at angle θ . The chord that passes through x perpendicularly to the radius is unique.



The corresponding sample space is

$$\Omega = \{(\theta, x) \mid \theta \in [0, 2\pi), x \in [0, 1]\}.$$

Because we choose the angle and radius uniformly and independently on their respective intervals, we see that for any $[a, b] \times [c, d] \subset [0, 2\pi) \times [0, 1]$ that

$$\mathbf{P}([a, b] \times [c, d]) = \frac{b - a}{2\pi} \cdot (d - c).$$

Define a random variable $X: \Omega \rightarrow \mathbb{R}$ by

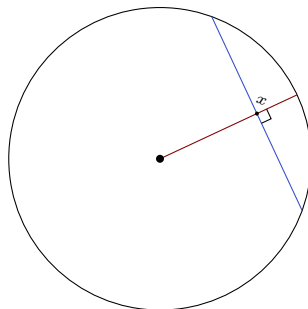
$$X((\theta, x)) = \text{the length of the chord corresponding to the pair } (\theta, x).$$

It turns out that $X(\theta, x) = 2\sqrt{1 - x^2}$. Now we see that

$$\mathbf{P}(X \geq \sqrt{3}) = \mathbf{P}([0, 2\pi] \times [0, 1/2]) = \frac{2\pi - 0}{2\pi} \cdot (1/2 - 0) = \frac{1}{2}.$$

7.2 Second solution

Now in order to choose a chord, we choose an arbitrary point inside the circle, and consider the chord perpendicular to the radius that is intersecting the with the point.



The corresponding sample space is

$$\Omega = \{(x, y) \mid x^2 + y^2 \leq 1\}.$$

Because we choose the coordinates uniformly, and independently, we see that for any $[a, b] \times [c, d] \subset \Omega$ that

$$\mathbf{P}([a, b] \times [c, d]) = \frac{(b-a)(d-c)}{\pi},$$

which is just the area of the rectangle divided by the area of the circle. Define a random variable $Y: \Omega \rightarrow \mathbb{R}$ by

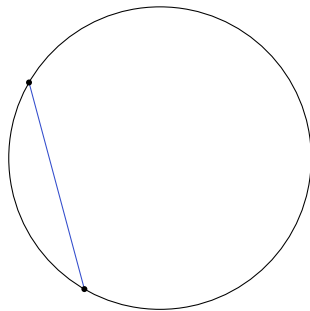
$$Y((x, y)) = \text{the length of the chord corresponding to the pair } (x, y).$$

It turns out that $Y(x, y) = 2\sqrt{1 - x^2 - y^2}$. Now the result we get is

$$\mathbf{P}(Y \geq \sqrt{3}) = \mathbf{P}\left(\{(x, y) \mid 2\sqrt{1 - x^2 - y^2} \geq \sqrt{3}\}\right) = \mathbf{P}\left(\{(x, y) \mid x^2 + y^2 \leq \frac{1}{4}\}\right) = \frac{\frac{\pi}{4}}{\pi} = \frac{1}{4}.$$

7.3 Third solution

Now we simply choose a chord by choosing two points on the circle, and connecting them.



The corresponding sample space is

$$\Omega = \{(\theta_1, \theta_2) \mid \theta_1, \theta_2 \in [0, 2\pi)\}.$$

Because we choose the angles uniformly, and independently, we see that for any $[a, b] \times [c, d] \subset \Omega$ that

$$\mathbf{P}([a, b] \times [c, d]) = \frac{b-a}{2\pi} \cdot \frac{d-c}{2\pi},$$

Define a random variable

$$Z((\theta_1, \theta_2)) = \text{the length of the chord corresponding to the pair } (x, y).$$

As you may expect, calculating Z directly this time requires tedious geometric and trigonometric calculations. After these calculations, we use similar methods to the ones shown in the first two solutions, and get

$$\mathbf{P}(Z \geq \sqrt{3}) = \frac{1}{3}.$$

7.4 Conclusion

The conclusion to the paradox is that probability questions only make sense once the sample space and measure are clearly specified. The reason there seems to be a paradox in the first place, is that the English expression “suppose a chord of the circle is chosen uniformly at random”, seems like it determines the problem uniquely, while in fact there are multiple ways to interpret it, as we saw in the different solutions we gave.

8 Multivariate random variables

As we saw in the beginning of Section 4, random variables encode information about our experiments. In some experiments, we may want to measure multiple results, so a natural way to do it is by defining multiple random variables $X_1, \dots, X_n: \Omega \rightarrow \mathbb{R}$, one for each result we want to measure. A natural question to ask may be, what is the probability that $X_1 = X_2$? In other words what is $\mathbf{P}(X_1 = X_2)$? Defining multiple random variables is indeed a good way to do measure multiple results, and as we will see, we call $\mathbf{X} = (X_1, \dots, X_n)$ random vectors, but as the following example shows, we must be careful.

Example 8.1. Suppose we roll two fair dices, and let $X, Y: \Omega \rightarrow \mathbb{R}$ be random variables with the distribution $\mathbf{P}_X(i) = \mathbf{P}_Y(i) = 1/6$ for all $i \in \{1, 2, 3, 4, 5, 6\}$. We now want to compute $\mathbf{P}(X = Y)$. It may seem easy to say, well, each variable describes the roll of a fair die, so the probability that $X = Y$ is exactly $1/6$. However, just like we saw in Section 7, it is very important to have a perfect correspondence between the experiment, and our mathematical description. In this case, if indeed X and Y were the results of rolling two fair dice, then we would have $\mathbf{P}(X = Y) = 1/6$. However, suppose that we had a different experiment, say, we roll one fair dice, and we write the same result twice, where $\tilde{X}$ is the first number we wrote, and $\tilde{Y}$ is the second number we wrote. In this case, when we write the experiment in a mathematical way, unfortunately, we have that $\tilde{X}, \tilde{Y}: \tilde{\Omega} \rightarrow \mathbb{R}$ are such that $\mathbf{P}_X = \mathbf{P}_{\tilde{X}}$, and $\mathbf{P}_Y = \mathbf{P}_{\tilde{Y}}$, but $\mathbf{P}(X = Y) = 1$. This shows that random variables encode more information than their *marginal*¹ probability distributions. Indeed, if we compare the sample space of both experiments, or consider the random variables as the functions that they are we could see that $X \neq Y \neq \tilde{X} = \tilde{Y}$, while $\mathbf{P}_X = \mathbf{P}_Y = \mathbf{P}_{\tilde{X}} = \mathbf{P}_{\tilde{Y}}$. This shows that the way we first described the experiment has multiple interpretations and does not correspond to our experiment exactly, unless we assume the reader *implicitly* defines the sample space and the random variables as functions to be exactly how we want them to be.

8.1 Multivariate random variables

Definition 8.1 (Multivariate random variable). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. The function $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$ is said to be a multivariate random variable if for all $a_i < b_i$ for $1 \leq i \leq N$ we have

$$\mathbf{X}^{-1} \left(\prod_{i=1}^N (a_i, b_i) \right) = \{ \omega \in \Omega \mid a_i < X_i(\omega) < b_i, \forall 1 \leq i \leq N \} \in \mathcal{F}.$$

A multivariate random variable is sometimes also called a random vector.

Remark 8.1. Notice that this definition also implies that $\mathbf{X}^{-1}(A) \in \mathcal{F}$ for all $A \in \mathfrak{B}(\mathbb{R}^n)$.

Proposition 8.1. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$. Then $\mathbf{X}$ is a multivariate random variable if and only if $X_k: \Omega \rightarrow \mathbb{R}$ is a random variable for all $1 \leq k \leq N$.

Proof. Assume $\mathbf{X}$ is a multivariate random variable and set some $1 \leq k \leq N$. Then for all $a < b$ we have

$$X_k^{-1}((a, b)) = \mathbf{X}^{-1} \left(\mathbb{R}^{k-1} \times (a, b) \times \mathbb{R}^{n-k} \right) \in \mathcal{F}.$$

¹When we consider the distribution of a random vector we sometimes use the phrasing “joint distribution”, and when we consider the random variables as separate from the vectors we sometimes use terms such as “marginal distribution”.

Now assume $X_1, \dots, X_N$ are random variables. Then for all $a_i < b_i$ for $1 \leq i \leq N$ we have

$$\mathbf{X}^{-1} \left(\prod_{i=1}^N (a_i, b_i) \right) = \bigcap_{i=1}^N \{ \omega \in \Omega \mid a_i < X_i(\omega) < b_i \} \in \mathcal{F},$$

where the last equality follows from the fact $\mathcal{F}$ is a σ -algebra, and $X_1, \dots, X_N$ are all random variables. $\square$

Definition 8.2 (Distribution of a multivariate random variable). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$ be a multivariate random variable. The (joint) distribution of $\mathbf{X}$, denoted $\mathbf{P}_{\mathbf{X}}$ is a probability function on the measurable space $(\mathbb{R}^N, \mathfrak{B}(\mathbb{R}^n))$ defined as:

$$\mathbf{P}_{\mathbf{X}}(A) = \mathbf{P}(\mathbf{X} \in A), \quad \forall A \in \mathfrak{B}(\mathbb{R}^N).$$

Remark 8.2. The function $\mathbf{P}_{X_k}$ denotes the distribution of the random variable X_k and is called the partial distribution of X_k . We see that the distribution $\mathbf{P}_{\mathbf{X}}$ of a random vector uniquely defines the partial distributions $\mathbf{P}_{X_1}, \dots, \mathbf{P}_{X_N}$ because

$$\mathbf{P}_{X_k} = \mathbf{P}_{\mathbf{X}} \left(\mathbb{R}^{k-1} \times A \times \mathbb{R}^{N-k} \right), \quad \forall A \in \mathfrak{B}.$$

As we saw in Example 8.1, the opposite is not true - the partial distributions do not determine the joint distribution uniquely.

Exercise 8.1. Suppose there are 3 red balls, 4 white balls, and 5 blue balls in a vase. Three balls are arbitrarily drawn out of the vase. Denote the number of red balls drawn by X , and the number of white balls drawn by Y . What is the distribution of the random vector (X, Y) ?

Solution. We see that

$$\Omega = \{ (i, j) \mid i, j \geq 0 \text{ and } i + j \leq 3 \},$$

where i denotes the number of red balls drawn, and j the number of white balls drawn. From a simple combinatorial arguments we get

$$\mathbf{P}_{(X,Y)}((i, j)) = \frac{\binom{3}{i} \binom{4}{j} \binom{5}{3-i-j}}{\binom{12}{3}}.$$

Definition 8.3 (Cumulative distribution function). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$ be a random vector. The cumulative distribution function of $\mathbf{X}$, denoted $F_{\mathbf{X}}$, is the function $F_{\mathbf{X}}: \mathbb{R}^N \rightarrow [0, 1]$ defined as

$$F_{\mathbf{X}}(a_1, \dots, a_N) = \mathbf{P}(X_i \leq a_i, \forall 1 \leq i \leq N) = \mathbf{P}_{\mathbf{X}} \left(\prod_{i=1}^N (-\infty, a_i] \right).$$

Proposition 8.2 (Properties of the CDF). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$ be a random vector. Then

- (1) The function $F_{\mathbf{X}}$ is increasing in each coordinate.
- (2) The function $F_{\mathbf{X}}$ is right-continuous in each coordinate.
- (3) Fix $a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_N \in \mathbb{R}$, then we have $\lim_{a_k \rightarrow -\infty} F_{\mathbf{X}}(a_1, a_2, \dots, a_N) = 0$.

(4) Denote $X^k = (X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_N)$ the random vector with the k th coordinate removed. Then for every choice $a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_N \in \mathbb{R}$ we have

$$\lim_{a_k \rightarrow \infty} F_{\mathbf{X}}(a_1, a_2, \dots, a_N) = F_{X^k}(a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_N).$$

In particular

$$\lim_{a_1, \dots, a_N \rightarrow \infty} F_{\mathbf{X}}(a_1, a_2, \dots, a_N) = 1.$$

Proof. The proofs are quite straightforward. It is important to remember that when trying to prove that an increasing function f converges to infinity, it suffices to prove that $f(a_n) \rightarrow \infty$ for any $a_n \rightarrow \infty$ you choose. This lemma is more useful in this course than other courses due to the nature of distribution functions (they are always increasing). $\square$

Remark 8.3. The following two theorems are generalizations of Theorem 4.2 and Theorem 4.3 respectively. They allow us to conclude that there exists a bijection between all random vectors' distributions and functions that satisfy the 4 properties from Theorem 8.2.

Theorem 8.3. Every cumulative distribution function corresponds to a unique distribution. In other words, if $(\Omega, \mathcal{F}, \mathbf{P})$ and $(\Omega', \mathcal{F}', \mathbf{P}')$ are two probability spaces, and $\mathbf{X}: \Omega \rightarrow \mathbb{R}^N$, $\mathbf{Y}: \Omega' \rightarrow \mathbb{R}^N$ are two random variable, then

$$F_{\mathbf{X}} = F_{\mathbf{Y}} \iff \mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{Y}}$$

Theorem 8.4. Let $F: \mathbb{R}^N \rightarrow [0, 1]$ be a function that satisfies all 4 properties in Theorem 8.2. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a random vector $\mathbf{X}: \Omega \rightarrow \mathbb{R}^N$ such that $F_{\mathbf{X}} = F$.

8.2 Theory of random vectors

Definition 8.4 (Discrete random vector). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$ be a random vector. Then $\mathbf{X}$ is said to be a discrete random vector if there exists a countable set $B \subset \mathbb{R}^N$ such that $\mathbf{P}(\mathbf{X} \in B) = 1$.

Proposition 8.5. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$ be a random vector. Then $\mathbf{X}$ is a discrete random vector if and only if X_k is a discrete random variable for all $1 \leq k \leq N$.

Proof. Assume that $\mathbf{X} = (X_1, \dots, X_N)$ is a discrete random vector. Then there exists a countable set $B \subset \mathbb{R}^N$ such that

$$\mathbf{P}_{\mathbf{X}}(B) = 1.$$

Let $1 \leq k \leq N$. Consider the set $B_k = \pi_k(B)$ where π_k is the projection on the k th coordinate. It is clear that B_k is countable and also

$$1 \geq \mathbf{P}(X_1 \in B_k) = \mathbf{P}\left((X_1, \dots, X_N) \in B_k \times \mathbb{R}^{N-1}\right) \geq \mathbf{P}(\mathbf{X} \in B) = 1,$$

where the last inequality follows from the fact that $B \subset B_k \times \mathbb{R}^{N-1}$. Thus we got that for all $1 \leq k \leq N$ that X_k is discrete. Now suppose that $X_1, \dots, X_N$ are discrete random variables. Then exist countable sets $B_1, \dots, B_N$ such that $\mathbf{P}(X_k \in B_k) = 1$ for all $1 \leq k \leq N$. The set $B = B_1 \times \dots \times B_N$ is countable and it is clear that $\mathbf{P}(\mathbf{X} \in B) = 1$ as wanted which completes the proof. $\square$

Remark 8.4. Notice that the above proposition proves that given a discrete random vector $\mathbf{X} = (A_i)_{i=1}^N$, for any subsequence $\{i_j\}_{j=1}^k$ of $\{i\}_{i=1}^N$ that $\mathbf{X} = (A_{i_j})_{j=1}^k$ is also a discrete random vector.

Definition 8.5 (Continuous random vector). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$ be a random vector. Then $\mathbf{X}$ is called a continuous random vector if $F_{\mathbf{X}}$ is a continuous function.

Remark 8.5. As in the case of a random variable, if $\mathbf{X}: \Omega \rightarrow \mathbb{R}^N$ is a continuous random vector, then $\mathbf{P}(\mathbf{X} = a) = 0$ for all $a \in \mathbb{R}^N$.

Remark 8.6. The full theory of continuous random vectors is outside the scope of this course and we will only focus on absolutely continuous random vectors.

Definition 8.6 (Absolutely continuous random vector). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$ be a random vector. Then $\mathbf{X}$ is said to be absolutely continuous if there exists an integrable function $f: \mathbb{R}^N \rightarrow [0, \infty)$ such that

$$\int_{\mathbb{R}^N} f(x) dx = 1,$$

and such that for all $a \in \mathbb{R}^N$ we have:

$$F_{\mathbf{X}}(a) = \mathbf{P}_{\mathbf{X}} \left(\prod_{i=1}^N (-\infty, a_i] \right) = \int_{\prod_{i=1}^N (-\infty, a_i]} f(y_1, \dots, y_N) d(y_1, \dots, y_N).$$

In this case, we call $f_{\mathbf{X}}$ the probability density function of $\mathbf{X}$, or the PDF of $\mathbf{X}$ for short.

Remark 8.7. Similarly to the theory of random variables, this definition implies that for any $A \subset \mathbb{R}^N$, such that $f \cdot \mathbb{1}_A$ is an integrable function, we get

$$\mathbf{P}_{\mathbf{X}}(A) = \int_A f(y_1, \dots, y_N) d(y_1, \dots, y_N).$$

Remark 8.8. Let (X, Y) be a continuous random vector with distribution function $F(x, y)$. Assume that F piecewise twice continuously differentiable. Then $f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y)$. Of course this can be generalized to n -dimensional random vectors as well.

Remark 8.9. Our main tool for dealing with integrals in $\mathbb{R}^n$ is called Fubini's theorem. We are going to use the version of the theorem that generalizes Fubini's theorem from analysis 2 to functions on $\mathbb{R}^n$.

Theorem 8.6 (Fubini's theorem). Let $f: A \times B \rightarrow \mathbb{R}$ be Riemann integrable, where $A, B \subset \mathbb{R}$ are closed sets. For all $x \in A$ define the function $f_x: B \rightarrow \mathbb{R}$ by $f_x(y) = f(x, y)$. If for all $x \in A$ the function f_x is Riemann integrable on B , then

$$\int_{A \times B} f(x, y) d(x, y) = \int_A \left(\int_B f_x(y) dy \right) dx.$$

Similarly, for each $y \in B$ we define the function $f^y: A \rightarrow \mathbb{R}$ by $f^y(x) = f(x, y)$. If for all $y \in B$ the function f^y is Riemann integrable on A , then

$$\int_{A \times B} f(x, y) d(x, y) = \int_B \left(\int_A f^y(x) dx \right) dy.$$

Exercise 8.2. Let (X, Y) be a random vector with a probability density function given by

$$f_{(X,Y)}(x, y) = \begin{cases} c(2x + y^2) & 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

- (1) Find the value of c .
- (2) Calculate $\mathbf{P}(X + Y \leq 1)$.

Solution.

- (1) In order to find c we can use the fact that $\int_{\mathbb{R}^2} f_{(X,Y)}(x,y)d(x,y) = 1$. From Fubini's theorem we get that

$$\begin{aligned} 1 &= \int_{\mathbb{R}^2} f_{(X,Y)}(x,y)d(x,y) = \int_0^1 \left(\int_0^2 c(2x + y^2)dy \right) dx = c \int_0^1 \left[2xy + \frac{y^3}{3} \right]_0^2 dx \\ &= c \int_0^1 4x + \frac{8}{3} dx = \frac{14}{3}c, \end{aligned}$$

which implies that $c = \frac{3}{14}$.

- (2) In order to find $\mathbf{P}(X + Y \leq 1)$, we need to understand that the function is only different than zero on the rectangle $R = [0, 1] \times [0, 2]$. This means that we need to find the integral of f_X on $T = R \cap \{(x, y) \mid x + y \leq 1\}$. Using Fubini's theorem we get

$$\begin{aligned} \mathbf{P}(X + Y \leq 1) &= \int_T f_{(X,Y)}(x,y)d(x,y) = \frac{3}{14} \int_0^1 \left(\int_0^{1-x} 2x + y^2 dy \right) dx \\ &= \frac{3}{14} \int_0^1 2x(1-x) + \frac{(1-x)^3}{3} dx = \frac{5}{56}. \end{aligned}$$

Theorem 8.7. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \dots, X_N): \Omega \rightarrow \mathbb{R}^N$ be an absolutely continuous random vector with a probability density function $f_{\mathbf{X}}: \mathbb{R}^N \rightarrow [0, \infty)$. Then, for all $1 \leq k \leq N$, the random variable X_k is absolutely continuous with a probability density function

$$f_{X_k}(a) = \int_{\mathbb{R}^{N-1}} f_{\mathbf{X}}(x_1, \dots, x_{k-1}, a, x_{k+1}, \dots, x_N) d(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_N).$$

Definition 8.7 (Volume). Let $A \subset \mathbb{R}^N$ be a set. We define the volume of A as such

$$\text{Vol}(A) = \int_{\mathbb{R}^N} \mathbb{1}_A(x) dx.$$

Definition 8.8 (Uniform distribution). Let $\mathbf{X}: \Omega \rightarrow \mathbb{R}^N$ be a random vector, and $B \subset \mathbb{R}^N$ an open or closed set with a well defined, finite volume. The random vector $\mathbf{X}$ is said to distribute uniformly on B , and we write $\mathbf{X} \sim U(B)$, if $\mathbf{X}$ is absolutely continuous with a probability density function

$$f_{\mathbf{X}}(x) = \begin{cases} \frac{1}{\text{Vol}(B)} & x \in B, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 8.9 (Independence). Let $X_1, \dots, X_N: \Omega \rightarrow \mathbb{R}$ be random variables. The variables $X_1, \dots, X_N$ are said to be independent if for all $A_1, \dots, A_N \in \mathfrak{B}$ we have

$$\mathbf{P}(X_1 \in A_1, \dots, X_N \in A_N) = \prod_{i=1}^N \mathbf{P}(X_i \in A_i).$$

Remark 8.10. Equivalently, we can write that $X_1, \dots, X_N$ are independent if for all Borel measurable sets $A_1, \dots, A_N$ we have

$$\mathbf{P} \left(\bigcap_{i=1}^N \{X_i \in A_i\} \right) = \prod_{i=1}^N \mathbf{P}(X_i \in A_i).$$

This means that $X_1, \dots, X_N$ are independent if and only if for all Borel measurable sets $A_1, \dots, A_N$ the events $(\{X_i \in A_i\})_{i=1}^N$ are independent. In terms of distributions, this means that

$$\mathbf{P}_{\mathbf{X}}(A_1 \times \dots \times A_N) = \prod_{i=1}^N \mathbf{P}_{X_i}(A_i).$$

Wait. This formula shows that when $X_1, \dots, X_N$ are independent variables, the partial distributions $\mathbf{P}_{X_i}$ do determine $\mathbf{P}_{\mathbf{X}}$ uniquely. If you forgot why this is so surprising, recall Example 8.1.

Proposition 8.8. Let $X_1, \dots, X_N: \Omega \rightarrow \mathbb{R}$ be random variables, and denote $\mathbf{X} = (X_1, \dots, X_N)$. Then

- (1) If $X_1, \dots, X_N$ are independent, then $F_{\mathbf{X}}(a_1, \dots, a_N) = \prod_{i=1}^N F_{X_i}(a_i)$ for all $a_1, \dots, a_N \in \mathbb{R}$.
- (2) If $X_1, \dots, X_N$ are independent, and $\mathbf{X}$ is absolutely continuous, then $f_{\mathbf{X}}(a_1, \dots, a_N) = \prod_{i=1}^N f_{X_i}(a_i)$ for all $a_1, \dots, a_N \in \mathbb{R}$.

Remark 8.11. In fact both (1) and (2) are if and only if statements.

Example 8.2. Let X and Y be two random variables so that $\mathbf{P}(X = i) = \mathbf{P}(Y = i) = 1/2$ for $i \in \{0, 1\}$. Then X and Y are independent if and only if $\mathbf{P}(X = i, Y = j) = 1/4$ for any $i, j \in \{0, 1\}$. Recall that once we know the partial distributions $\mathbf{P}_X$ and $\mathbf{P}_Y$, and that X and Y are independent, then we can compute $\mathbf{P}_{(X,Y)}$.

Exercise 8.3. Let $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$, and suppose that X_1 and X_2 are independent. Find $\mathbf{P}(X_2 > X_1)$.

Solution. We know that X_i is absolutely continuous with

$$f_{X_i}(x) = \begin{cases} \lambda_i e^{-\lambda_i x} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

for all $i \in \{1, 2\}$. Since X_1 and X_2 are independent, by Theorem 8.8 we conclude that

$$f_{(X_1, X_2)}(x, y) = f_{X_1}(x)f_{X_2}(y) = \begin{cases} \lambda_1 \lambda_2 e^{\lambda_1 \lambda_2 xy} & x \geq 0 \text{ and } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, we have that

$$\mathbf{P}(X_2 > X_1) = \int_{\{(x,y)|y>x\}} f_{(X_1, X_2)}(x, y) d(x, y).$$

However, we can see that the only part of $\{(x, y) \mid y > x\}$ where $f_{(X_1, X_2)}(x, y)$ does not vanish is $\{(x, y) \mid y > x > 0\}$, so we can use Fubini's theorem to conclude that

$$\begin{aligned} \mathbf{P}(X_2 > X_1) &= \int_{\{(x,y)|y>x\}} f_{(X_1, X_2)}(x, y) d(x, y) = \int_0^\infty \left(\int_x^\infty \lambda_1 \lambda_2 e^{\lambda_1 \lambda_2 xy} dy \right) dx \\ &= \int_0^\infty \lambda_1 e^{-\lambda_1 x} \left(\int_x^\infty \lambda_2 e^{-\lambda_2 y} dy \right) dx = \int_0^\infty \lambda_1 e^{-\lambda_1 x} \left[-e^{-\lambda_2 y} \right]_x^\infty dx \\ &= \lambda_1 \int_0^\infty e^{\lambda_1 \lambda_2 x^2} dx = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \end{aligned}$$

Definition 8.10 (Measure). Let (X, Σ) be a measurable space. A function $\mu: \Sigma \rightarrow [0, \infty]$ is called a measure if the following conditions hold

- (1) $\mu(\emptyset) = 0$;
- (2) For all countable collections $\{E_k\}_{k=1}^{\infty}$ of pairwise disjoint sets in Σ we have

$$\mu\left(\bigsqcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k).$$

Remark 8.12. Condition (2) in Definition 8.10 is called σ -additivity.

Remark 8.13. Note that if $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space, then $\mathbf{P}$ is a measure on $(\Omega, \mathcal{F})$. Conversely, any probability function $\mathbf{P}: \Omega \rightarrow [0, 1]$ is a measure on $(\Omega, \mathcal{F})$, so that $\mu(\Omega) = 1$. That is why in measure theory, a probability function is called a probability measure.

Definition 8.11 (Atom). Let (X, Σ) be a measurable space, and let μ be a measure on that space. A set $A \in \Sigma$ is called an atom, if the following conditions holds:

- (1) $\mu(A) > 0$.
- (2) For any measurable set $B \subset A$, either $\mu(B) = 0$, or $\mu(A \setminus B) = 0$.

Remark 8.14. The intuition for Definition 8.11 is that it cannot be split into smaller measurable sets (with measure greater than zero), just like an atom. This definition is not going to be very important from now on, but I still included it because it appears from time to time when talking about discrete random vectors.

Remark 8.15. Note that in the case of a distribution $\mathbf{P}_{\mathbf{X}}: \mathfrak{B}(\mathbb{R}^N) \rightarrow \mathbb{R}$, its atoms must be singletons.

Remark 8.16. Similarly to Theorem 4.8, we can decompose every random vector $\mathbf{X}: \Omega \rightarrow \mathbb{R}^N$ to the sum of a discrete and a continuous random vector. The proof of this decomposition is outside the scope of the course, but we will still give the formal statement.

Theorem 8.9. Let $\mathbf{X}: \Omega \rightarrow \mathbb{R}^N$ be a random vector, and let $\{a_i\}_{i=1}^{\infty} \subset \mathbb{R}^N$ be the atoms of the distribution $\mathbf{P}_{\mathbf{X}}$. Then there exists a discrete and continuous distribution functions $F_{\mathbf{X}}^d$ and $F_{\mathbf{X}}^c$ respectively, so that $F_{\mathbf{X}} = \alpha F_{\mathbf{X}}^d + (1 - \alpha) F_{\mathbf{X}}^c$.

Remark 8.17. We can generalize Theorem 4.9 to random vectors. If $\mathbf{X}: \Omega \rightarrow \mathbb{R}^N$ is a random vector, and $h: \mathbb{R}^N \rightarrow \mathbb{R}^M$ is a measurable function, then $\mathbf{Y} = h(\mathbf{X})$ is a random vector. We also want to generalize Theorem 4.11 and describe $F_{\mathbf{Y}}$ with $F_{\mathbf{X}}$. However, instead of talking about the general case, in this course, we will only talk about what happens in the cases of a few simple functions, all of which are of the form $h: \mathbb{R}^n \rightarrow \mathbb{R}$.

Example 8.3 (Maximum of two variables). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be random variables, and define $Z: \Omega \rightarrow \mathbb{R}$ by $Z = \max\{X, Y\}$. Note that $h: (x, y) \mapsto \max(x, y)$ is continuous and therefore it is measurable, so $Z = h(X, Y)$ is a random variable. Our goal is to compute F_Z in terms of the distributions of X and Y . We get that

$$F_Z(a) = \mathbf{P}(Z \leq a) = \mathbf{P}(X \leq a, Y \leq a) = F_{(X, Y)}(a, a).$$

For example if $X \sim U([0, 1])$, $Y \sim U([0, 1])$, and X and Y are independent, then

$$F_Z(a) = F_{(X, Y)}(a, a) = F_X(a)F_Y(a) = \begin{cases} 0 & a < 0, \\ a^2 & 0 \leq a \leq 1, \\ 1 & a > 1, \end{cases}$$

where the second transition follows from the fact X and Y are independent.

Example 8.4 (Minimum of two variables). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be random variables, and define $Z: \Omega \rightarrow \mathbb{R}$ by $Z = \min\{X, Y\}$. Note that $h: (x, y) \mapsto \min(x, y)$ is continuous and therefore it is measurable, so $Z = h(X, Y)$ is a random variable. Our goal is to compute F_Z in terms of the distributions of X and Y . We get that

$$\begin{aligned} F_Z(a) &= \mathbf{P}(Z \leq a) = \mathbf{P}(X \leq a \text{ or } Y \leq a) \\ &= \mathbf{P}(X \leq a) + \mathbf{P}(Y \leq a) - \mathbf{P}(X \leq a, Y \leq a) = F_X(a) + F_Y(a) - F_{(X,Y)}(a, a). \end{aligned}$$

For example if $X \sim U([0, 1])$, $Y \sim U([0, 1])$, and X and Y are independent, then

$$F_Z(a) = F_X(a) + F_Y(a) - F_{(X,Y)}(a, a) = F_X(a) + F_Y(a) - F_X(a)F_Y(a) = \begin{cases} 0 & a < 0, \\ 2a - a^2 & 0 \leq a \leq 1, \\ 1 & a > 0, \end{cases}$$

where the second transition follows from the fact X and Y are independent.

Example 8.5 (Sum of two independent variables). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be independent random variables, and define $Z = X + Y$. Note that $h: (x, y) \mapsto x + y$ is continuous and therefore it is measurable, so $Z = h(X, Y)$ is a random variable. Our goal is to compute F_Z in terms of the distributions of X and Y . First, assume that X and Y are discrete with support $\{a_i\}_{i=1}^{\infty}$ and $\{b_i\}_{i=1}^{\infty}$ respectively. We get that

$$F_Z(a) = \mathbf{P}(X + Y \leq a) = \sum_{i=1}^{\infty} \mathbf{P}(X + Y \leq a, X = a_i) = \sum_{i=1}^{\infty} \mathbf{P}(Y \leq a - a_i, X = a_i).$$

Since X and Y are independent we conclude that

$$F_Z(a) = \sum_{i=1}^{\infty} \mathbf{P}(Y \leq a - a_i) \mathbf{P}(X = a_i) = \sum_{i=1}^{\infty} F_Y(a - a_i) \mathbf{P}_X(a_i).$$

Now assume that X and Y are absolutely continuous, so $f_{(X,Y)} = f_X(x)f_Y(y)$, and thus

$$\begin{aligned} F_{X+Y}(a) &= \mathbf{P}(X + Y \leq a) = \int_{\{(x,y)|x+y \leq a\}} f_{(X,Y)} d(x, y) = \int_{\{(x,y)|x+y \leq a\}} f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{a-x} f_X(x) f_Y(y) dy \right) dx = \int_{-\infty}^{\infty} f_X(x) \left(\int_{-\infty}^{a-x} f_Y(y) dy \right) dx \\ &= \int_{-\infty}^{\infty} f_X(x) F_Y(a - x) dx. \end{aligned}$$

Note that in both cases we got that $F_{X+Y}(a)$ is a sum (or integral) over one function valued at x , and another function valued at $a - x$. This type of expression is fundamental in all kinds of fields of mathematics, most prominently in Fourier analysis, and it is called a *convolution*.

Definition 8.12 (Convolution). Let $f, g: \mathbb{R} \rightarrow [0, \infty)$. The convolution of f and g , denoted $f * g: \mathbb{R} \rightarrow [0, \infty)$, is the function

$$(f * g)(a) := \int_{-\infty}^{\infty} f(x) g(a - x) dx.$$

Remark 8.18. We can also define convolutions for $f, g: \mathbb{R} \rightarrow \mathbb{R}$ such that $|f|$ and $|g|$ are integrable. These type of functions are called $\mathcal{L}^1(\mathbb{R})$ functions.

Remark 8.19. The convolution operator is well defined, commutative ($f * g = g * f$), and associative ($(f * g) * h = f * (g * h)$).

Remark 8.20. The computations in Example 8.5 show that for random variables $X, Y: \Omega \rightarrow \mathbb{R}$ we have $F_{X+Y} = f_X * F_Y = f_Y * F_X$, and $f_{X+Y} = f_X * f_Y$. Also note that the fact $f_{X+Y} = f_X * f_Y$, shows that $X + Y$ is absolutely continuous.

Example 8.6. Let $X, Y \sim U([0, 1])$ be independent random variables. Then

$$f_{X+Y}(a) = (f_X * f_Y)(a) = \int_{-\infty}^{\infty} f_X(x) f_Y(x-a) dx = \int_0^1 f_Y(x-a) dx = |[0, 1] \cap [a-1, a]|.$$

By a matter of calculation

$$f_{X+Y}(a) = \begin{cases} 0 & a \leq 0, \\ a & 0 \leq a \leq 1, \\ 2-a & 1 \leq a \leq 2, \\ 0 & a \geq 2. \end{cases}$$

Exercise 8.4 (The Rayleigh distribution). Let $(X, Y): \Omega \rightarrow \mathbb{R}^2$ be an absolutely continuous random vector with $f_{(X,Y)}(x, y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$. Denote $R = \sqrt{X^2 + Y^2}$, and compute $f_R(r)$

Solution. We have that

$$F_R(r) = \mathbf{P}(\sqrt{X^2 + Y^2} \leq r) = \mathbf{P}(X^2 + Y^2 \leq r^2) = \int_{\{(x,y) | |(x,y)| \leq r\}} f_{(X,Y)}(x, y) d(x, y).$$

By the change of variables theorem

$$F_R(r) = \int_0^{2\pi} d\theta \int_0^r \frac{1}{2\pi} e^{-\frac{r^2}{2}} \cdot r dr = \int_0^r e^{-\frac{r^2}{2}} \cdot r dr = \begin{cases} 1 - e^{-r^2/2} & r \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$f_R(r) = F'_R(r) = \begin{cases} r e^{-r^2/2} & r \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

This distribution is called the Rayleigh distribution with parameter $\sigma = 1$.

9 Expectation of random vectors, covariance and correlation

9.1 Expectation

Remark 9.1 (Expectation of random vectors). A natural extension of expectation to random vectors is to set

$$\mathbf{E}[\mathbf{X}] = \int_{\mathbb{R}^N} \mathbf{y} f_{\mathbf{X}}(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^N} (y_1, \dots, y_N) f_{\mathbf{X}}(y_1, \dots, y_N) d(y_1, \dots, y_N).$$

Recall that an integral of vectors is defined as the vector of the integrals

$$\mathbf{E}[\mathbf{X}] = \left(\int_{\mathbb{R}^N} y_1 f_{\mathbf{X}}(y_1, \dots, y_N) d(y_1, \dots, y_N), \dots, \int_{\mathbb{R}^N} y_N f_{\mathbf{X}}(y_1, \dots, y_N) d(y_1, \dots, y_N) \right).$$

Now we can consider the first coordinate and using Fubini's theorem we get

$$\begin{aligned} \int_{\mathbb{R}^N} y_1 f_{\mathbf{X}}(y_1, \dots, y_N) d(y_1, \dots, y_N) &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}^{N-1}} y_1 f_{\mathbf{X}}(y_1, \dots, y_N) d(y_2, \dots, y_N) \right] dy_1 \\ &= \int_{\mathbb{R}} y_1 \left[\int_{\mathbb{R}^{N-1}} f_{\mathbf{X}}(y_1, \dots, y_N) d(y_2, \dots, y_N) \right] dy_1 \\ &= \int_{\mathbb{R}} y_1 f_{X_1}(y_1) dy_1 \\ &= \mathbf{E}[X_1], \end{aligned}$$

where the second to last equality follows from Theorem 8.7. This leads us to the following definition.

Definition 9.1 (Expectation). Let $\mathbf{X}: \Omega \rightarrow \mathbb{R}^N$ be a random vector such that $\mathbf{E}[X_i]$ is well defined and finite for all $1 \leq i \leq N$. Then we define

$$\mathbf{E}[\mathbf{X}] = (\mathbf{E}[X_1], \dots, \mathbf{E}[X_N]).$$

Proposition 9.1. Let $\mathbf{X}: \Omega \rightarrow \mathbb{R}^N$ be a random vector, and let $h: \mathbb{R}^N \rightarrow \mathbb{R}$ be a measurable function such that the random variable $|h(\mathbf{X})|$ has a finite expectation. Then if $\mathbf{X}$ is absolutely continuous with a probability density function $f_{\mathbf{X}}$, then

$$\mathbf{E}[h(\mathbf{X})] = \int_{\mathbb{R}^N} h(\mathbf{y}) f_{\mathbf{X}}(\mathbf{y}) d\mathbf{y}.$$

And if $\mathbf{X}$ is a discrete random vector with $\text{supp}(\mathbf{P}_{\mathbf{X}}) = (\mathbf{a}^i)_{i \geq 1}$, then

$$\mathbf{E}[h(\mathbf{X})] = \sum_{i=1}^{\infty} h(\mathbf{a}^i) \mathbf{P}_{\mathbf{X}}(\mathbf{a}^i).$$

Remark 9.2. Note that Theorem 9.1 is a generalization of the law of the unconscious statistician we saw in Theorem 5.2 and Theorem 5.3.

Corollary 9.2 (Linearity of expectation). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two random variables and $a, b \in \mathbb{R}$. Then

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y].$$

Proof. The idea is using Theorem 9.1 with the function $h(x, y) = ax + by$. □

Corollary 9.3. Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two random variables. If X and Y are independent then

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y].$$

This implies,

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

However, the converse is not always true.

Proof. The idea for the first part is using Theorem 9.1 with the function $h(x, y) = xy$. The second part follows by applying the first part and linearity of expectation. $\square$

Remark 9.3. As mentioned in the proof of Theorem 9.3, the equation $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ follows algebraically without any conditions once $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$. Even if X and Y are not independent.

Example 9.1. Let $X \sim U([-1, 1])$ and $Y = X^2$. It is clear that X and Y are not independent, but by the law of the unconscious statistician we get

$$\mathbf{E}[XY] = \mathbf{E}[X^3] = \int_{-1}^1 \frac{1}{2}x^3 dx = 0,$$

but since $\mathbf{E}[X] = 0$, we have that

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y],$$

and thus immediately

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

This shows that the converse to Theorem 9.3 is not always true.

9.2 Covariance and correlation

Definition 9.2 (Uncorrelatedness). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two random vectors. Then X and Y are said to be uncorrelated if $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

Remark 9.4. If X, Y are two independent random variables, then by Theorem 9.3 they are uncorrelated, but as we saw in Example 9.1, two uncorrelated random variables are not necessarily independent.

Remark 9.5. In general, from the definition of variance and the linearity of expectation, we *always* have

$$\text{Var}(X + Y) = \text{Var}(X) + 2\mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] + \text{Var}(Y).$$

Another equality we always have is

$$\mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

Intuitively, the value of $\mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$ represents how much X and Y are correlated, and it is called the covariance of X and Y . It is easier to see the intuition behind it when considering the expression $\mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$. It tells us that if X and Y tend to deviate above or below their expectation together, their correlation is positive, and the stronger the mutual deviation is, the higher the covariance, which is, their correlatedness. If X deviates above $\mathbf{E}[X]$ when Y deviates below $\mathbf{E}[Y]$, then X and Y are still correlated but the sign of the covariance will be negative. If there is not consistent relationship between their deviations from their means, then the positive and negative products cancel out on average, and their covariance will be zero, that is, they are uncorrelated.

Definition 9.3 (Covariance). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two random variables. The covariance of X and Y , denoted $\text{Cov}(X, Y)$ is defined as

$$\text{Cov}(X, Y) := \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

Remark 9.6. From the definition of covariance we get the following convenient formula:

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

Additionally, the variables X, Y are uncorrelated if and only if $\text{Cov}(X, Y) = 0$. This is because $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ also implies $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

Definition 9.4 (Correlation). The correlation of X, Y is denoted $\rho_{X,Y}$, and defined by

$$\rho_{X,Y} := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}.$$

Remark 9.7. The correlation $\rho_{X,Y}$ is a normalization of the covariance. Its value is bounded between -1 and 1 , where $\rho_{X,Y} = \pm 1$ implies a perfect positive/negative linear correlation respectively. We will show this soon.

Exercise 9.1. Suppose we conducted $n \geq 2$ independent experiments with three possible outcomes: 1, 2, and 3. Denote the probabilities of the first outcome by p_1 , the second outcome by p_2 , and the third outcome by $p_3 = 1 - p_1 - p_2$. Denote the number of times we got 1 by X , and the number of times we got 2 by Y . Compute $\text{Cov}(X, Y)$ and $\rho_{X,Y}$.

Solution. Since the values of X or Y have an inverse correlation, we expect $\text{Cov}(X, Y)$ to be negative, and we also expect $\text{Cov}(X, Y)$ to grow as $n \rightarrow \infty$. As we will see, this is indeed the case.

Since $X \sim \text{Bin}(n, p_1)$ and $Y \sim \text{Bin}(n, p_2)$ we get that $\mathbf{E}[X] = np_1$ and $\mathbf{E}[Y] = np_2$. Note that

$$X = \sum_{i=1}^n \mathbb{1}_{\{\text{the result is 1}\}} \quad \text{and} \quad Y = \sum_{i=1}^n \mathbb{1}_{\{\text{the result is 2}\}},$$

where the indicators are also random variables (make sure you understand how). Now by linearity of expectation we get

$$\begin{aligned} \mathbf{E}[XY] &= \mathbf{E} \left[\left(\sum_{i=1}^n \mathbb{1}_{\{\text{the result is 1}\}} \right) \left(\sum_{i=1}^n \mathbb{1}_{\{\text{the result is 2}\}} \right) \right] \\ &= \sum_{i,j=1}^n \mathbf{E} \left[\mathbb{1}_{\{\text{the result is 1}\}} \mathbb{1}_{\{\text{the result is 2}\}} \right]. \end{aligned}$$

We see that when $i = j$, we get $\mathbb{1}_{\{\text{the result is 1}\}} \mathbb{1}_{\{\text{the result is 2}\}} = 0$, and if $i \neq j$, then the events are independent, in which case we get

$$\begin{aligned} \mathbf{E} \left[\mathbb{1}_{\{\text{the } i\text{th result is 1}\}} \mathbb{1}_{\{\text{the } j\text{th result is 2}\}} \right] &= \mathbf{E} \left[\mathbb{1}_{\{\text{the } i\text{th result is 1}\}} \right] \mathbf{E} \left[\mathbb{1}_{\{\text{the } j\text{th result is 2}\}} \right] \\ &= \mathbf{P}(\text{the } i\text{th result is 1}) \mathbf{P}(\text{the } j\text{th result is 2}) = p_1 p_2. \end{aligned}$$

Therefore, since there are $n(n-1)$ pairs (i, j) such that $i \neq j$ we conclude that

$$\mathbf{E}[XY] = n(n-1)p_1 p_2.$$

We can now compute that

$$\text{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = n(n-1)p_1 p_2 - np_1 \cdot np_2 = -np_1 p_2.$$

Finally, because the variance of a binomial random variable is $np(1-p)$ we have that

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = \frac{-np_1p_2}{\sqrt{np_1(1-p_1)np_2(1-p_2)}} = -\sqrt{\frac{p_1p_2}{(1-p_1)(1-p_2)}}.$$

Proposition 9.4 (Properties of covariance). *Let $X, X_1, X_2, Y: \Omega \rightarrow \mathbb{R}$ be random variables, and let $a, b \in \mathbb{R}$. Then*

- (1) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
- (2) $\text{Cov}(aX, Y) = a \text{Cov}(X, Y)$.
- (3) $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$.
- (4) $\text{Cov}(X, X) \geq 0$.
- (5) $\text{Cov}(X, X) = 0 \iff \text{Var}(X) = 0 \iff \mathbf{P}(X = \mathbf{E}[X]) = 1$ (X is constant).
- (6) $\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$.

Proof. The proofs for all of these properties follow from the definitions of variance, covariance, or the linearity of expectation. Make sure these properties coincide with your intuition of correlatedness. $\square$

Remark 9.8 (The space $\mathcal{L}^2$). Notice that all random variables on a certain probability space $(\Omega, \mathcal{F}, \mathbf{P})$ form a vector space. We can also notice that the only thing that is preventing the map $(X, Y) \mapsto \text{Cov}(X, Y)$ from defining an inner product, is that there exist random variables $X \neq 0$ such that $\text{Cov}(X, X) = 0$. Notably, all the constant random variables that are different than 0 satisfy $\text{Cov}(X, X) = 0$. Luckily, there is a way to fix this, by requiring the expectation of all random variables to be 0. Indeed, if we define the space

$$\mathcal{L}^2(\Omega, \mathcal{F}, \mathbf{P}) = \{X: \Omega \rightarrow \mathbb{R} \mid X \text{ is a random variable and } \mathbf{E}[X] = 0 \text{ and } \text{Var}(X) < \infty\},$$

then the space $(\mathcal{L}^2(\Omega, \mathcal{F}, \mathbf{P}), \langle \cdot, \cdot \rangle)$, where $\langle \cdot, \cdot \rangle = \text{Cov}(\cdot, \cdot)$ is an inner product space, and moreover, it can be shown that it is a Hilbert space. By the Cauchy–Schwarz inequality we get that

$$|\text{Cov}(X, Y)| = |\langle X, Y \rangle| \leq \sqrt{\langle X, X \rangle} \cdot \sqrt{\langle Y, Y \rangle} = \sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}$$

for all $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbf{P})$. In fact, this inequality holds for any two random variables. If $\text{Var}(X) = \infty$ or $\text{Var}(Y) = \infty$ it holds very trivially, so we may assume that $\text{Var}(X), \text{Var}(Y) < \infty$. In this case, we know that $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ are both vectors in $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbf{P})$ and thus

$$\begin{aligned} |\text{Cov}(X, Y)| &= |\text{Cov}(X - \mathbf{E}[X], Y - \mathbf{E}[Y])| \leq \sqrt{\text{Var}(X - \mathbf{E}[X])} \cdot \sqrt{\text{Var}(Y - \mathbf{E}[Y])} \\ &= \sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}. \end{aligned}$$

Remark 9.9. The last equation in Remark 9.8 implies that the correlation of two random variables gives values in the interval $[-1, 1]$. That is, for any $X, Y: \Omega \rightarrow \mathbb{R}$ we have $-1 \leq \rho_{X,Y} \leq 1$. This justifies Remark 9.7.

Proposition 9.5. *Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two random variables. Then X and Y are independent if and only if for all measurable $f, g: \mathbb{R} \rightarrow \mathbb{R}$ such that $f(X), g(Y), f(X)g(Y)$ are integrable we have*

$$\mathbf{E}[f(X)g(Y)] = \mathbf{E}[f(X)]\mathbf{E}[g(Y)].$$

9.3 Indicators and expectation

As we saw in Exercise 9.1, the linearity of expectation allows us to simplify many calculations, especially if we use indicator-like random variables like in Exercise 9.1.

Example 9.2. Let $X \sim \text{Bin}(n, p)$. Since all the experiments are independent we get that $X = \sum_{i=1}^n X_i$ for each $X_i \sim \text{Bin}(1, p)$. Each X_i is an indicator-like random variable which gives 1 whenever the i th experiment is a success, and 0 otherwise. From the linearity of expectation we have

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{E}[X_i] = \sum_{i=1}^n p = np.$$

Since the variables $X_1, \dots, X_n$ are independent, we can also easily calculate the variance:

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) = \sum_{i=1}^n p(1-p) = np(1-p).$$

Note that we could have also computed this directly from the definition, however these calculations are a lot more simple and elegant.

Remark 9.10. This method of using indicators to find the expectation is very useful and is sometimes called the “method of indicators” for calculating expectation. One of the things that make it an incredibly strong method, is that it works even when the indicators are not independent, as can be seen in the following example.

Exercise 9.2. Let $\sigma: [n] \rightarrow [n]$ be a random permutation from the group of all permutations of n elements S_n . Denote the number of fixed points of σ by X . Find $\mathbf{E}[X]$.

Solution. First we notice that

$$X = \sum_{i=1}^n \mathbb{1}_{i \text{ is a fixed point}}$$

and thus

$$\begin{aligned} \mathbf{E}[X] &= \mathbf{E}\left[\sum_{i=1}^n \mathbb{1}_{i \text{ is a fixed point}}\right] = \sum_{i=1}^n \mathbf{E}[\mathbb{1}_{i \text{ is a fixed point}}] = \sum_{i=1}^n \mathbf{P}(i \text{ is a fixed point}) \\ &= \sum_{i=1}^n \frac{(n-1)!}{n!} = \sum_{i=1}^n \frac{1}{n} = 1. \end{aligned}$$

Definition 9.5 (Covariance matrix). Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector, so that the expectation and variance of each X_i is finite. Then we define the covariance matrix of $\mathbf{X}$ by

$$C(\mathbf{X}) := \begin{pmatrix} \text{Cov}(X_1, X_1) & \cdots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \cdots & \text{Cov}(X_n, X_n) \end{pmatrix}.$$

Remark 9.11. By the properties of covariance, $C(\mathbf{X})$ is a symmetric matrix, and its diagonal is $\text{diag}(C(\mathbf{X})) = (\text{Var}(X_1), \dots, \text{Var}(X_n))$.

9.4 Cauchy distribution

Definition 9.6 (Cauchy distribution). An absolutely continuous random variable X is said to have Cauchy distribution if its distribution function is given by $f(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$.

Remark 9.12. We have that $\int \frac{1}{1+x^2} dx = \arctan(x)$ and also that $\int_0^\infty \frac{1}{1+x^2} dx = \frac{\pi}{2}$. From this it follows that if X is a Cauchy random variable then:

$$F_X(x) = \int_{-\infty}^x f(y) dy = \frac{1}{\pi} \arctan(y)|_{-\infty}^x = \frac{1}{\pi} \left(\arctan(x) + \frac{\pi}{2} \right),$$

for $-\infty \leq x \leq \infty$. This has the important implication that X has no defined expectation or variance since the integral $\int_{-\infty}^\infty x f(x) dx$ does not exist.

Remark 9.13 (One-sided Cauchy distribution). We say that an absolutely continuous random variable X has one-sided Cauchy distribution if its density is $f(x) = \frac{2}{\pi} \cdot \frac{1}{1+x^2}$ for $x \geq 0$ and zero otherwise. When we want to be precise we may call the distribution in Definition 9.6 the two-sided Cauchy distribution.

Exercise 9.3. Let X and Y be i.i.d. random variables so that they both distribute according to $N(0, \sigma^2)$. Show that the random variable $W := \frac{Y}{|X|}$ has the Cauchy distribution.

Solution. Denote $A_w = \{(x, y) \in \mathbb{R}^2 \mid y/|x| \leq w\}$, so that:

$$F_W(w) = \mathbf{P}\left(\frac{Y}{|X|} \leq w\right) = \mathbf{P}((X, Y) \in A_w) = \int_{A_w} f(x, y) dx dy,$$

where $f(x, y)$ is the joint density of X and Y , and since they are independent we have that:

$$f(x, y) = f(x) \cdot f(y) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \cdot \frac{e^{-\frac{y^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} = \frac{e^{-\frac{(x+y)^2}{2\sigma^2}}}{2\pi\sigma}.$$

Computing the integral with calculus 2 tools gives that:

$$F_W(w) = \frac{1}{\pi\sigma^2} \int_0^\infty dx \int_{-\infty}^{wx} \frac{e^{-\frac{(x+y)^2}{2\sigma^2}}}{2\pi\sigma} dy,$$

and by differentiation:

$$f_W(w) = F'_W(w) = \dots = \frac{1}{\pi} \cdot \frac{1}{1+w^2},$$

which completes the exercise.

10 Conditional probability distribution and expectation

10.1 The discrete case

Definition 10.1 (Conditional distribution, discrete case). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two discrete random variables, and let $a \in \mathbb{R}$ such that $\mathbf{P}_X(a) > 0$. The conditional distribution of Y given that $X = a$ is denoted $\mathbf{P}_{Y|X}(\cdot | a): \mathfrak{B} \rightarrow \mathbb{R}$ and defined as

$$\mathbf{P}_{Y|X}(B | a) := \mathbf{P}(Y \in B | X = a) = \frac{\mathbf{P}(Y \in B, X = a)}{\mathbf{P}(X = a)}$$

Proposition 10.1 (Properties of conditional distribution). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two discrete random variables, and let $a \in \mathbb{R}$ such that $\mathbf{P}_X(a) > 0$. Then

- (1) $\mathbf{P}_{Y|X}(\cdot | a)$ is a probability function.
- (2) $\mathbf{P}_Y(b) = \sum_a \mathbf{P}_X(a) \mathbf{P}_{Y|X}(b | a)$ for all $b \in \mathbb{R}$.
- (3) If $\mathbf{P}_Y(b) > 0$, then

$$\mathbf{P}_{X|Y}(a | b) = \frac{\mathbf{P}_{Y|X}(b | a) \mathbf{P}_X(a)}{\mathbf{P}_Y(b)}.$$

Proof. The proofs of all of these properties follow from the definition of conditional probability, the law of total probability, and Bayes' theorem. $\square$

Example 10.1. Let $\mathbf{X} = (X_1, X_2)$ be a discrete random vector given by

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	0.4	0.1
$X_2 = 1$	0.2	0.3

In this case by definition we have that

$$\mathbf{P}_{X_1|X_2}(0 | 0) = \frac{\mathbf{P}_{(X_1, X_2)}((0, 0))}{\mathbf{P}_{X_2}(0)} = \frac{0.4}{0.5} = 0.8 \quad \mathbf{P}_{X_1|X_2}(1 | 0) = \frac{\mathbf{P}_{(X_1, X_2)}((1, 0))}{\mathbf{P}_{X_2}(0)} = \frac{0.1}{0.5} = 0.2.$$

Exercise 10.1. Let X be the number of successes in N independent Bernoulli experiments with a probability of success p , and suppose that $N \sim \text{Poi}(\lambda)$. Find the distribution of X .

Solution. Given $N = n$, we know that $X \sim \text{Bin}(n, p)$. Therefore, for $k \geq 0$ we have

$$\begin{aligned} \mathbf{P}_X(k) &= \sum_{n=0}^{\infty} \mathbf{P}_N(n) \mathbf{P}_{X|N}(k | n) = \sum_{n=k}^{\infty} \mathbf{P}_N(n) \mathbf{P}_{X|N}(k | n) \\ &= \sum_{n=k}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{e^{-\lambda}}{k!} p^k \lambda^k \sum_{n=k}^{\infty} \frac{\lambda^{n-k}}{(n-k)!} (1-p)^{n-k} \\ &= \frac{e^{-\lambda}}{k!} p^k \lambda^k \sum_{n=0}^{\infty} \frac{\lambda^n}{(n)!} (1-p)^n = \frac{e^{-\lambda}}{k!} p^k \lambda^k e^{\lambda(1-p)} = e^{-\lambda p} \frac{(\lambda p)^k}{k!}. \end{aligned}$$

This means that $X \sim \text{Poi}(\lambda p)$.

Remark 10.1 (Conditional distribution of multiple variables). Given discrete random variables $X, Y, Z: \Omega \rightarrow \mathbb{R}$, if we assume that $\mathbf{P}_Z(c) > 0$, we could also define that

$$\mathbf{P}_{(X,Y)|Z}(B | c) = \mathbf{P}((X, Y) \in B | Z = c),$$

for any event $B \subset \mathbb{R}^2$. And if we assume $\mathbf{P}_{(X,Y)}(a, b) > 0$, we could also define

$$\mathbf{P}_{Z|(X,Y)}(C | (a, b)) = \mathbf{P}(Z \in C | (X, Y) = (a, b)),$$

for any event $A \subset \mathbb{R}$. Of course, we can generalize this definition for any finite amount of random variables.

Definition 10.2 (Conditional expectation, given an event). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two discrete random variables, and let $a \in \mathbb{R}$ be such that $\mathbf{P}_X(a) > 0$. The condition expectation of Y given that $X = a$ is the expectation of the random variable with distribution $\mathbf{P}_{Y|X}(\cdot | a)$. In other words,

$$\mathbf{E}[Y | X = a] = \sum_b b \mathbf{P}_{Y|X}(b | a).$$

Exercise 10.2. Let $X, Y \sim \text{Geo}(p)$ be independent random variables, and denote $Z = X + Y$. Find $\mathbf{E}[X | Z = m]$ for $m \geq 2$.

Solution. Recall that the range of geometric random variables is $\mathbb{N}$, so $Z = X + Y \geq 2$, and moreover, given that $X + Y = m$ we know that X only has atoms in $\{1, \dots, m-1\}$. Thus,

$$\begin{aligned} \mathbf{P}_{X|Z}(k | m) &= \frac{\mathbf{P}(X = k, Y = m - k)}{\mathbf{P}(Z = m)} = \frac{\mathbf{P}(X = k, Y = m - k)}{\sum_{j=1}^{m-1} \mathbf{P}(X = j, Y = m - j)} \\ &= \frac{\mathbf{P}(X = k) \mathbf{P}(Y = m - k)}{\sum_{j=1}^{m-1} \mathbf{P}(X = j) \mathbf{P}(Y = m - j)}, \end{aligned}$$

where the last inequality is from the fact X and Y are independent. Now by plugging in the values we conclude that

$$\mathbf{P}_{X|Z}(k | m) = \frac{p(1-p)^{k-1} \cdot p(1-p)^{m-k-1}}{\sum_{j=1}^{m-1} p(1-p)^{j-1} \cdot p(1-p)^{m-j-1}} = \frac{p^2(1-p)^{m-2}}{\sum_{j=1}^{m-1} p^2(1-p)^{m-2}} = \frac{1}{m-1}.$$

This means that given $Z = m$, the variable X distributes uniformly on $\{1, \dots, m-1\}$, and thus

$$\mathbf{E}[X | Z = m] = \sum_{k=1}^{m-1} k \mathbf{P}_{X|Z}(k | m) = \frac{1}{m-1} \sum_{k=1}^{m-1} k = \frac{1}{m-1} \cdot \frac{(m-1)m}{2} = \frac{m}{2}.$$

Remark 10.2. So far, we have only talked about computing the distribution or expectation of a variable X given that an event of the form $\{Y = b\}$ has occurred. For example, in the previous exercise (Exercise 10.2), we saw that $\mathbf{E}[X | Z = m] = m/2$. This means that $\mathbf{E}[X | Z = m]$ is a function in m , and indeed, any conditional expectation $\mathbf{E}[X | Z = m]$ will always be some function of m (possibly a constant function), that is, a function from $\{a | \mathbf{P}_Z(a) > 0\}$ to $\mathbb{R}$. Since $X: \Omega \rightarrow \mathbb{R}$ is a function from Ω to $\{a | \mathbf{P}_Z(a) > 0\}$, this allows us to take our theory a step further, and define a new random variable by composing these two functions. It turns out that this composition is a random variable.

Definition 10.3 (Conditional expectation, given a variable). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two discrete random variables. Suppose that $\text{supp}(\mathbf{P}_X) = \{a_i\}_{i \geq 1}$. The conditional expectation of Y given X is a random variable $\mathbf{E}[Y | X]: \Omega \rightarrow \mathbb{R}$ defined by

$$\mathbf{E}[Y | X](\omega) = \begin{cases} \mathbf{E}[Y | X = a_1], & \text{if } X(\omega) = a_1 \\ \mathbf{E}[Y | X = a_2], & \text{if } X(\omega) = a_2 \\ \mathbf{E}[Y | X = a_3], & \text{if } X(\omega) = a_3 \\ \vdots & \vdots \end{cases}$$

Remark 10.3. We can equivalently write

$$\mathbf{E}[Y | X](\omega) = \sum_{i=1}^{\infty} \mathbf{E}[Y | X = a_i] \cdot \mathbb{1}_{X=a_i}(\omega).$$

Proposition 10.2 (Properties of conditional expectation). *Let $X, Y: \Omega \rightarrow \mathbb{R}$ be two discrete random variables. Then*

- (1) $\mathbf{E}[X | Y]$ is a random variable.
- (2) If $\mathbf{E}[|Y|] < \infty$, then $\mathbf{E}[\mathbf{E}[Y | X]] = \mathbf{E}[Y]$.
- (3) If $h: \mathbb{R} \rightarrow \mathbb{R}$ is measurable such that $\mathbf{E}[|h(X)Y|] < \infty$, then $\mathbf{E}[h(X)Y | X] = h(X)\mathbf{E}[Y | X]$, as random variables.

Remark 10.4. The properties of conditional expectation shown in Theorem 10.2 are very handy when computing the expectancy of random variables. Especially property (2), which is sometimes called the “Law of Total Expectation”.

Exercise 10.3. Let X be the number of successes in N Bernoulli trials with probability of success $p > 0$, and suppose that $N \sim \text{Poi}(\lambda)$. Compute $\mathbf{E}[X]$ and $\mathbf{E}[X/N]$.

Solution. Recall that after a tedious computation from Exercise 10.1 we found that $X \sim \text{Poi}(\lambda p)$, so we can conclude that $\mathbf{E}[X] = \lambda p$. However, before we move on to finding the conditional expectation $\mathbf{E}[X | N]$, we will see a different way of finding $\mathbf{E}[X]$ by using conditional expectation. Since $X \sim \text{Bin}(N, p)$, and N is positive, we get that X is nonnegative, and we assume it has finite expectation (well, we know this from earlier), by the law of total expectation (and that $\mathbf{E}[X | N] = Np$) we get

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X | N]] = \mathbf{E}[Np] = p\mathbf{E}[N] = p\lambda.$$

The fact that $\mathbf{E}[X | N] = Np$ follows from $\mathbf{E}[X | N = n] = np$. We can now move on to compute $\mathbf{E}[X/N]$. By using properties (2) and (3) (with the function $h(x) = 1/x$) we get

$$\mathbf{E}[X/N] = \mathbf{E}[\mathbf{E}[X/N | N]] = \mathbf{E}\left[\frac{1}{N}\mathbf{E}[X | N]\right] = \mathbf{E}\left[\frac{Np}{N}\right] = \mathbf{E}[p] = p.$$

10.2 The absolutely continuous case

Recall that in the discrete case we have defined

$$\mathbf{P}_{Y|X}(B | a) := \mathbf{P}(Y \in B | X = a) = \frac{\mathbf{P}(Y \in B, X = a)}{\mathbf{P}(X = a)}.$$

However, in the absolutely continuous case, we have that $\mathbf{P}(X = a) = 0$, so this is impossible. Instead, we will define conditional distribution through the probability density functions.

Definition 10.4 (Conditional probability density function). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be absolutely continuous random variables with a joint probability density function $f_{(X,Y)}$, and partial density functions f_X and f_Y respectively. For any $a \in \mathbb{R}$ so that $f_X(a) > 0$, we define f_Y given $X = a$, denoted $f_{Y|X}(\cdot | a): \mathbb{R} \rightarrow [0, \infty)$ by

$$f_{Y|X}(b | a) := \frac{f_{(X,Y)}(a, b)}{f_X(a)}.$$

Proposition 10.3. *The function $f_{Y|X}(\cdot | a)$ is a probability function of a random variable. That is, $f_{Y|X}(\cdot | a)$ is nonnegative, integrable, and $\int_{-\infty}^{\infty} f_{Y|X}(y | a) dy = 1$.*

Definition 10.5 (Conditional distribution, absolutely continuous case). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be absolutely continuous random variables, and let $Y | X = a$ denote the random variable associated with $f_{Y|X}(\cdot | a)$. We define the distribution of $Y | X = a$ by

$$\mathbf{P}_{Y|X=a}(A) := \int_A f_{Y|X}(y | a) dy.$$

Remark 10.5. We sometimes write $\mathbf{P}(Y \in A | X = a)$ instead of $\mathbf{P}_{Y|X=a}(A)$ for convenience.

Proposition 10.4 (Law of total probability). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be absolutely continuous random variables. Then

$$f_Y(y) = \int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y | x) dx.$$

Proposition 10.5 (Bayes' theorem). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be absolutely continuous random variables, and let $x, y \in \mathbb{R}$ be such that $f_X(x) > 0$ and $f_Y(y) > 0$. Then

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

Proof. The proofs follow immediately from the definitions. □

Definition 10.6 (Conditional expectation, absolutely continuous variables). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be absolutely continuous random variables with a joint probability density function $f_{(X,Y)}$, and partial density functions f_X and f_Y respectively. For all $x \in \mathbb{R}$ such that $f_X(x) > 0$ we define the conditional expectation of Y given that $X = x$, to be the expectation of the random variable $Y | X = x$, or in other words,

$$\mathbf{E}[Y | X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y | x) dy.$$

Remark 10.6. Just like in the discrete case, we can use conditional expectation given an event, to define the random variable of conditional expectation given a variable $\mathbf{E}[Y | X]$. The formal definition of this random variable requires tools that are outside the scope of this course, but even though it is not the most formal definition, we could define $\mathbf{E}[Y | X]: \Omega \rightarrow \mathbb{R}$ by

$$\mathbf{E}[Y | X](\omega) = \mathbf{E}[Y | X = x] \Big|_{x=X(\omega)}.$$

Moreover, just like the discrete case, we have the law of total expectation in the absolutely continuous case as well.

Proposition 10.6 (Properties of conditional expectation). Let $X, Y: \Omega \rightarrow \mathbb{R}$ be absolutely continuous random variables with finite expectation. Then

- (1) $\mathbf{E}[\mathbf{E}[Y | X]] = \mathbf{E}[Y]$.
- (2) For any measurable function $h: \mathbb{R} \rightarrow \mathbb{R}$ such that $\mathbf{E}[|h(X)Y|] < \infty$ we have $\mathbf{E}[h(X)Y | X] = h(X)\mathbf{E}[Y | X]$.

Exercise 10.4. Let X, Y be two absolutely continuous random variables with probability density function

$$F_{(X,Y)}(x, y) = \begin{cases} \frac{1}{y} e^{-\frac{x}{y}-y} & x, y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Compute $f_{X|Y}$ and $\mathbf{P}(X > 1 | Y = y)$.

Solution. First notice that when $y \geq 0$ we get

$$\int_{-\infty}^{\infty} f_{(X,Y)}(x, y) dx = \frac{1}{y} e^{-y} \int_{-\infty}^{\infty} e^{-\frac{x}{y}} dx = e^{-y}.$$

This implies that

$$f_Y(y) = \begin{cases} e^{-y} & y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

This means that $Y \sim \text{Exp}(1)$. Now by definition of the conditional probability density function, for any fixed $y \geq 0$ we get

$$f_{X|Y}(x | y) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)} = \begin{cases} \frac{1}{y} e^{-\frac{x}{y}} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

This means that $X | Y = y \sim \text{Exp}(1/y)$. We can now compute that

$$\mathbf{P}(X > 1 | Y = y) = \int_1^{\infty} f_{X|Y}(x | y) dx = \int_1^{\infty} \frac{1}{y} e^{-\frac{x}{y}} dx = e^{-1/y}.$$

Exercise 10.5. Let $X \sim \text{Exp}(1)$, and suppose that $Y | X = x \sim U([x, x + 1])$. Find $f_Y(y)$ and $\mathbf{E}[Y]$.

Solution. Note that by definition

$$f_{Y|X}(y | x) = \mathbb{1}_{[x, x+1]}(y) = \begin{cases} 1 & x \leq y \leq x + 1, \\ 0 & \text{otherwise.} \end{cases}$$

This implies that

$$f_{(X,Y)}(x, y) = f_{Y|X}(y | x) f_X(x) = \begin{cases} e^{-x} & 0 \leq x \leq y \leq x + 1, \\ 0 & \text{otherwise.} \end{cases}$$

Finally,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y) dx = \begin{cases} \int_{y-1}^y e^{-x} dx & y \geq 1 \\ \int_0^y e^{-x} dx & 0 \leq y \leq 1 \\ 0 & y < 0 \end{cases} = \begin{cases} e^{-(y-1)} - e^{-y} & y \geq 1, \\ 1 - e^{-y} & 0 \leq y \leq 1, \\ 0 & y < 0. \end{cases}$$

We can now compute $\mathbf{E}[Y]$ in two ways. First, by definition we have

$$\mathbf{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_0^1 y (1 - e^{-y}) dy + \int_1^{\infty} y (e^{-(y-1)} - e^{-y}) dy = \frac{3}{2}.$$

The second way, is by using conditional expectation. By the law of total expectation

$$\mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y | X]] = \mathbf{E}\left[\frac{X + (X + 1)}{2}\right] = \mathbf{E}\left[X + \frac{1}{2}\right] = \mathbf{E}[X] + \frac{1}{2} = 1 + \frac{1}{2} = \frac{3}{2},$$

where the second step follows from the fact that $\mathbf{E}[X | Y]$ distributes uniformly on $[X, X + 1]$.

11 The Borel–Cantelli lemmas

11.1 Set theoretic preparation

Definition 11.1 (Limit superior). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\{A_j\}_{j=1}^{\infty}$ be a sequence of events in $\mathcal{F}$. We define

$$\limsup_{n \rightarrow \infty} A_n := \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n = \{\omega \in A_n \text{ infinitely often}\} = \{\omega \in A_n \text{ i.o.}\}.$$

Definition 11.2 (Limit inferior). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\{A_j\}_{j=1}^{\infty}$ be a sequence of events in $\mathcal{F}$. We define

$$\liminf_{n \rightarrow \infty} A_n := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k = \{\omega \in \bigcap_{k=n}^{\infty} A_n \text{ eventually}\}.$$

Remark 11.1. It is clear that $\liminf_{n \rightarrow \infty} A_n \subseteq \limsup_{n \rightarrow \infty} A_n$.

Definition 11.3 (Limit of sets). Let $\{A_n\}_{n=1}^{\infty}$ be an event sequence. Then A_n is said to have a limit if

$$\liminf_{n \rightarrow \infty} A_n = \limsup_{n \rightarrow \infty} A_n.$$

In this case, we denote

$$\lim_{n \rightarrow \infty} A_n := \liminf_{n \rightarrow \infty} A_n = \limsup_{n \rightarrow \infty} A_n.$$

Example 11.1. It is clear that if $\{A_n\}_{n=1}^{\infty}$ is an increasing sequence of events then $\lim_{n \rightarrow \infty} A_n = \bigcup_{n=1}^{\infty} A_n$, and when it is a decreasing sequence then $\lim_{n \rightarrow \infty} A_n = \bigcap_{n=1}^{\infty} A_n$.

Example 11.2. Let $\Omega = \mathbb{N}$, and define $A_k := \{k^j \mid j \in \mathbb{N}\}$. We get that $\lim_{n \rightarrow \infty} A_n = \{1\}$.

Example 11.3. Let $\Omega = \mathbb{N}$, and define

$$A_n := \begin{cases} 2\mathbb{N} & n \text{ is even,} \\ 2\mathbb{N} + 1 & n \text{ is odd.} \end{cases}$$

In this case, it is clear that $\limsup_{n \rightarrow \infty} A_n = \mathbb{N}$, and $\liminf_{n \rightarrow \infty} A_n = \emptyset$.

Theorem 11.1 (Continuity of the probability function, revisited). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\{A_n\}_{n=1}^{\infty}$ be a sequence of events in $\mathcal{F}$ such that the limit $\lim_{n \rightarrow \infty} A_n$ exists. Then,*

$$\lim_{n \rightarrow \infty} \mathbf{P}(A_n) = \mathbf{P}\left(\lim_{n \rightarrow \infty} A_n\right).$$

Remark 11.2. In order to prove Theorem 11.1, we need the following lemma.

11.2 Fatou's lemma

Lemma 11.2 (Fatou's lemma). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\{A_n\}_{n=1}^{\infty}$ be a sequence of events in $\mathcal{F}$. Then*

$$(1) \quad \mathbf{P}\left(\liminf_{n \rightarrow \infty} A_n\right) \leq \liminf_{n \rightarrow \infty} \mathbf{P}(A_n).$$

$$(2) \quad \mathbf{P}\left(\limsup_{n \rightarrow \infty} A_n\right) \geq \limsup_{n \rightarrow \infty} \mathbf{P}(A_n).$$

Proof. We will only prove (1), since the proof of (2) is symmetrical. Recall that

$$\liminf_{n \rightarrow \infty} A_n = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n,$$

and define $G_k := \bigcap_{n=k}^{\infty} A_n$. Notice that $\{G_k\}_{k=1}^{\infty}$ is an increasing sequence of events, and thus

$$\lim_{k \rightarrow \infty} G_k = \bigcup_{k=1}^{\infty} G_k = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n = \liminf_{n \rightarrow \infty} A_n.$$

Thus, by Theorem 1.6 we get that

$$\mathbf{P}(\liminf_{n \rightarrow \infty} A_n) = \mathbf{P}\left(\bigcup_{k=1}^{\infty} G_k\right) = \lim_{k \rightarrow \infty} \mathbf{P}(G_k) = \lim_{k \rightarrow \infty} \mathbf{P}\left(\bigcap_{n=k}^{\infty} A_n\right).$$

Note that from monotonicity of the probability function, for all $m \geq k$ we have that

$$\mathbf{P}\left(\bigcap_{n=k}^{\infty} A_n\right) \leq \mathbf{P}(A_m),$$

and therefore

$$\mathbf{P}\left(\bigcap_{n=k}^{\infty} A_n\right) \leq \inf_{m \geq k} \mathbf{P}(A_m).$$

We can now use this inequality and get that

$$\mathbf{P}(\liminf_{n \rightarrow \infty} A_n) = \lim_{k \rightarrow \infty} \mathbf{P}\left(\bigcap_{n=k}^{\infty} A_n\right) \leq \lim_{k \rightarrow \infty} \inf_{m \geq k} \mathbf{P}(A_m) = \liminf_{k \rightarrow \infty} \mathbf{P}(A_k),$$

which completes the proof. $\square$

Remark 11.3. The intuition for Fatou’s lemma is not as complex as it may seem. While the full explanation would simply turn into a verbal version of the proof, we can see that if our task were to find a set with smaller probability than $\liminf \mathbf{P}(A_n)$, it is very natural to derive the definition of $\liminf A_n$, and similarly it is very intuitive to derive $\liminf A_n$ when being asked to find a set with bigger probability than $\liminf \mathbf{P}(A_n)$ (of course we disregard trivial cases like choosing Ω or $\emptyset$).

Proof of Theorem 11.1. Suppose that $\lim_{n \rightarrow \infty} A_n$ exists. In this case we get from Fatou’s lemma that

$$\limsup_{n \rightarrow \infty} \mathbf{P}(A_n) \leq \mathbf{P}\left(\limsup_{n \rightarrow \infty} A_n\right) = \mathbf{P}\left(\lim_{n \rightarrow \infty} A_n\right) = \mathbf{P}\left(\liminf_{n \rightarrow \infty} A_n\right) \leq \liminf_{n \rightarrow \infty} \mathbf{P}(A_n).$$

Since $\{\mathbf{P}(A_n)\}_{n=1}^{\infty}$ is a sequence of numbers such that

$$\limsup_{n \rightarrow \infty} \mathbf{P}(A_n) \leq \liminf_{n \rightarrow \infty} \mathbf{P}(A_n),$$

the limit $\lim_{n \rightarrow \infty} \mathbf{P}(A_n)$ must exist, and from the squeeze theorem we get that

$$\lim_{n \rightarrow \infty} \mathbf{P}(A_n) = \mathbf{P}\left(\lim_{n \rightarrow \infty} A_n\right),$$

which completes the proof. $\square$

11.3 The Borel–Cantelli lemmas

The lemmas of Borel and Cantelli discuss the probability of the event $\limsup_{n \rightarrow \infty} A_n$. Note that since

$$\left(\limsup_{n \rightarrow \infty} A_n \right)^c = \left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n \right)^c = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n^c = \liminf_{n \rightarrow \infty} A_n^c,$$

any result on $\limsup_{n \rightarrow \infty} A_n$ can be translated and applied to $\liminf_{n \rightarrow \infty} A_n^c$.

Lemma 11.3 (First Borel–Cantelli lemma). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\{A_n\}_{n=1}^{\infty}$ be a sequence of events in $\mathcal{F}$ such that $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$. Then $\mathbf{P}(\limsup_{n \rightarrow \infty} A_n) = 0$.*

Proof. For $k \geq 1$ we define the event

$$G_k := \bigcup_{n=k}^{\infty} A_n.$$

It is clear that the sequence $\{G_k\}_{k=1}^{\infty}$ is decreasing. We also notice that

$$\bigcap_{k=1}^{\infty} G_k = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n = \limsup_{n \rightarrow \infty} A_n.$$

This implies that

$$\mathbf{P}\left(\limsup_{n \rightarrow \infty} A_n\right) = \mathbf{P}\left(\bigcap_{k=1}^{\infty} G_k\right) = \lim_{k \rightarrow \infty} \mathbf{P}(G_k).$$

By monotonicity of the probability function we get

$$\mathbf{P}(G_k) = \mathbf{P}\left(\bigcup_{n=k}^{\infty} A_n\right) \leq \sum_{n=k}^{\infty} \mathbf{P}(A_n).$$

Finally, we conclude that

$$0 \leq \mathbf{P}\left(\limsup_{n \rightarrow \infty} A_n\right) = \lim_{k \rightarrow \infty} \mathbf{P}(G_k) \leq \lim_{k \rightarrow \infty} \sum_{n=k}^{\infty} \mathbf{P}(A_n) = 0,$$

where the last transition follows from the assumption that $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$. This implies that $\mathbf{P}(\limsup_{n \rightarrow \infty} A_n) = 0$, which completes the proof. $\square$

Remark 11.4. The intuition for the first Borel–Cantelli lemma, is that if we know that $\lim_{k \rightarrow \infty} \sum_{n=k}^{\infty} \mathbf{P}(A_n) = 0$, then the measure (the probability) of the event with elements that appear infinitely often in $\{A_n\}_{n=1}^{\infty}$ cannot be greater than zero.

Example 11.4. Let $X_n \sim \text{Bin}(1, 1/n^2)$, and define $A_n := \{X_n \text{ is success}\}$. Since $\mathbf{P}(A_n) = 1/n^2$ and

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty,$$

the first lemma of Borel–Cantelli implies that $\mathbf{P}(\limsup_{n \rightarrow \infty} A_n) = 0$. Note that

$$\limsup_{n \rightarrow \infty} A_n = \{\omega \in \Omega \mid \text{we succeeded infinitely many times in } \omega\},$$

which means that the probability of succeeding infinitely many times is zero.

Remark 11.5. Note that there are examples where $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$, but $\mathbf{P}(\limsup A_n) = 0$. That is, the first Borel–Cantelli lemma is not an if and only if statement.

Example 11.5. Let $\Omega = [0, 1]$, let $\mathbf{P}$ be uniform on $[0, 1]$, and let $A_n = [0, 1/n]$. By definition of uniform probability we get $\mathbf{P}(A_n) = 1/n$, and therefore

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty,$$

however, $\limsup_{n \rightarrow \infty} A_n = \{0\}$, and therefore

$$\mathbf{P}\left(\limsup_{n \rightarrow \infty} A_n\right) = \mathbf{P}(\{0\}) = 0.$$

Remark 11.6. The second Borel–Cantelli lemma tells us that if $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$, and the events $A_1, \dots, A_n$ are independent, then we can conclude that $\mathbf{P}(\limsup_{n \rightarrow \infty} A_n) = 1$, and in particular, that $\mathbf{P}(\limsup_{n \rightarrow \infty} A_n) > 0$.

Lemma 11.4 (Second Borel–Cantelli lemma). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\{A_n\}_{n=1}^{\infty}$ be a sequence of independent events in $\mathcal{F}$, such that $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$. Then we have $\mathbf{P}(\limsup_{n \rightarrow \infty} A_n) = 1$.*

Proof. Recall that $(\limsup_{n \rightarrow \infty} A_n)^c = \liminf_{n \rightarrow \infty} A_n^c$, so it suffices to prove the equality $\mathbf{P}(\liminf_{n \rightarrow \infty} A_n^c) = 0$. Since we also have the inequality:

$$\mathbf{P}(\liminf_{n \rightarrow \infty} A_n^c) = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n^c \leq \sum_{k=1}^{\infty} \mathbf{P}\left(\bigcap_{n=k}^{\infty} A_n^c\right),$$

it suffices to show that $\mathbf{P}(\bigcap_{n=k}^{\infty} A_n^c) = 0$ for all $k \geq 1$. Let $k \geq 1$. Since $\{A_n\}_{n=1}^{\infty}$ are independent events, the events $\{A_n^c\}_{n=1}^{\infty}$ are also independent, so for any $M \geq k$ we get

$$\mathbf{P}\left(\bigcap_{n=k}^M A_n^c\right) = \prod_{n=k}^M \mathbf{P}(A_n^c) = \prod_{n=k}^M (1 - \mathbf{P}(A_n)) \leq \prod_{n=k}^M e^{-\mathbf{P}(A_n)} = \exp\left(-\sum_{n=k}^M \mathbf{P}(A_n)\right),$$

where the inequality follows from the known inequality $1 - x \leq e^{-x}$ (for $x \geq 0$). By taking the limits $M \rightarrow \infty$ on both sides, and by continuity of the functions we get

$$\mathbf{P}\left(\bigcap_{n=k}^{\infty} A_n^c\right) \leq 0,$$

which completes the proof. □

12 The law of large numbers

12.1 Convergence of random variables

Definition 12.1 (Convergence in probability). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X, X_n (n \in \mathbb{N}): \Omega \rightarrow \mathbb{R}$ be random variables. The sequence X_n is said to converge in probability to X if for all $\epsilon > 0$ we have

$$\lim_{n \rightarrow \infty} \mathbf{P}(|X_n - X| \geq \epsilon) = \lim_{n \rightarrow \infty} \mathbf{P}(\{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| \geq \epsilon\}) = 0.$$

In this case we denote $X_n \xrightarrow{P} X$.

Remark 12.1. In the context of measure theory, convergence in probability is called convergence in measure.

Definition 12.2 (Convergence in $\mathcal{L}^p$). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X, X_n (n \in \mathbb{N}): \Omega \rightarrow \mathbb{R}$ be random variables. Given $p \geq 1$, the sequence X_n is said to converge in $\mathcal{L}^p$ to X if

$$\lim_{n \rightarrow \infty} \mathbf{E}[|X_n - X|^p] = 0.$$

In this case we denote $X_n \xrightarrow{\mathcal{L}^p} X$. We sometimes say that X_n converges to X in the r th mean.

Remark 12.2. Usually we will only talk about convergence in $\mathcal{L}^1$ or $\mathcal{L}^2$. In these cases we say that X_n converges to X in mean, or in quadratic mean respectively.

Definition 12.3 (Convergence in distribution). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X, X_n (n \in \mathbb{N}): \Omega \rightarrow \mathbb{R}$ be random variables. We say that X_n converge in distribution to X (or converge weakly to X) if

$$\lim_{n \rightarrow \infty} F_{X_n}(a) = F_X(a),$$

for all $a \in \mathbb{R}$ where F_X is continuous. In this case we denote $X_n \xrightarrow{\mathcal{L}} X$, or $X_n \xrightarrow{d} X$, or $X_n \xrightarrow{w} X$, or $X_n \xrightarrow{\mathcal{D}} X$, or $X_n \Rightarrow X$.

Definition 12.4 (Almost surely convergence). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X, X_n (n \in \mathbb{N}): \Omega \rightarrow \mathbb{R}$ be random variables. We say that X_n converge almost surely to X (or converge strongly to X) if

$$\mathbf{P}\left(\left\{\omega \in \Omega \mid \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

In this case we denote $X_n \xrightarrow{\text{a.s.}} X$.

Example 12.1. Let $\Omega = [0, 1]$, let $\mathbf{P}$ be the uniform distribution function on Ω , and for all $n \geq 1$ define $X_n(\omega) = n^{-1}\omega$. It can be shown that $X_n \sim U([0, 1/n])$, or equivalently,

$$F_{X_n}(a) = \begin{cases} 0 & a < 0, \\ na & a \in [0, 1/n], \\ 1 & a > 1/n. \end{cases}$$

Define $X: \Omega \rightarrow \mathbb{R}$ by $X(\omega) = 0$. We will show that $X_n \rightarrow X$ in all four types of convergence we have introduced above.

$(X_n \xrightarrow{d} X)$ In order to show convergence in distribution, we first note that $F_X(a) = 0$ for $a < 0$, and $F_X(a) = 1$ otherwise. We also note that F_X is continuous in $\mathbb{R} \setminus \{0\}$, so we need to show that $\lim_{n \rightarrow \infty} F_{X_n}(a) = F_X(a)$ for all $a \in \mathbb{R} \setminus \{0\}$, and very easily one can verify that this is the case.

$(X_n \xrightarrow{p} X)$ We need to show that for all $\epsilon > 0$ we have

$$\lim_{n \rightarrow \infty} \mathbf{P} \left(\left\{ \omega \in \Omega \mid |X_n(\omega) - X(\omega)| \geq \epsilon \right\} \right) = 0.$$

Indeed, for all $n > 1/\epsilon$, we have

$$\mathbf{P} \left(\left\{ \omega \in \Omega \mid |X_n(\omega) - X(\omega)| \geq \epsilon \right\} \right) = \mathbf{P} \left(\left\{ \omega \in \Omega \mid \omega > n\epsilon \right\} \right) = \mathbf{P}(\emptyset) = 0,$$

so indeed $X_n \xrightarrow{p} X$.

$(X_n \xrightarrow{\mathcal{L}^p} X)$ Since $X_n \sim U([0, 1/n])$ and $X \equiv 0$ we get that for all $p \geq 1$ we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{E}[|X_n - X|^p] &= \lim_{n \rightarrow \infty} \mathbf{E}[X_n^p] = \lim_{n \rightarrow \infty} \int_0^{1/n} ny^p dy \\ &= \lim_{n \rightarrow \infty} \frac{n}{(p+1)n^{p+1}} = \lim_{n \rightarrow \infty} \frac{1}{(p+1)n^p} = 0. \end{aligned}$$

$(X_n \xrightarrow{\text{a.s.}} X)$ We have that

$$\left\{ \omega \in \Omega \mid \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \right\} = \left\{ \omega \in [0, 1] \mid \lim_{n \rightarrow \infty} \frac{\omega}{n} = 0 \right\} = [0, 1],$$

so

$$\mathbf{P} \left(\left\{ \omega \in \Omega \mid \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \right\} \right) = \mathbf{P}([0, 1]) = 1,$$

which shows that $X_n \xrightarrow{\text{a.s.}} X$.

Remark 12.3. We have the following implications between types of convergence:

$$\begin{array}{ccccc} (X_n \xrightarrow{\mathcal{L}^p} X) & & & & \\ p \geq q \geq 1 \downarrow & & & & \\ (X_n \xrightarrow{\mathcal{L}^q} X) & \longrightarrow & (X_n \xrightarrow{p} X) & \longleftarrow & (X_n \xrightarrow{\text{a.s.}} X) \\ & & \downarrow & & \\ & & (X_n \xrightarrow{d} X) & & \end{array}$$

However, any implication outside the ones inferred from this diagram is wrong in general. We will not provide a proof for all of these implications right now, but we will prove some of the directions when we discuss the law of large numbers.

Remark 12.4 (Generalizing convergence in distribution). Note that $X_n \xrightarrow{d} X$ if and only if $\lim_{n \rightarrow \infty} F_{X_n}(a) = F_X(a)$ for all $x \in \mathbb{R}$ where F_X is continuous. Note that this definition only depends on the distribution function of $X, X_n (n \in \mathbb{N})$, and not on the random variables themselves. This allows us to generalize this notion of convergence even for random variables that are not in the same probability space.

Definition 12.5 (Convergence of distribution functions). Let $F, F_n (n \in \mathbb{N}): \mathbb{R} \rightarrow \mathbb{R}$ be distribution functions. We say that F_n converges to F if $\lim_{n \rightarrow \infty} F_{X_n}(a) = F_X(a)$ for all $x \in \mathbb{R}$ where F_X is continuous.

Definition 12.6 (Convergence in distribution, generalized). Let $X_n (n \in \mathbb{N}): \Omega_n \rightarrow \mathbb{R}$ be random variables, and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. We say that X_n converges in distribution to X if F_n converges to F in the sense of Definition 12.5.

Remark 12.5. This generalized sense of convergence in distribution justifies writing $X_n \xrightarrow{d} \text{Exp}(\lambda)$, or $X_n \xrightarrow{d} \text{Bin}(\lambda)$.

Example 12.2. We have seen in Theorem 4.4 that

$$\lim_{n \rightarrow \infty} \mathbf{P}_X(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

This implies that

$$\lim_{n \rightarrow \infty} F_{X_n}(a) = \sum_{k=0}^{\lfloor a \rfloor} e^{-\lambda} \frac{\lambda^k}{k!},$$

so in other words $\text{Bin}(n, \lambda/n) \xrightarrow{d} \text{Poi}(\lambda)$.

Example 12.3. Let $X_n \sim \frac{1}{n} \text{Geo}(\lambda n^{-1})$. We see that

$$1 - F_{X_n}(a) = \mathbf{P}(X_n > a) = \mathbf{P}(\text{Geo}(\lambda n^{-1}) > na) = \left(1 - \frac{\lambda}{n}\right)^{\lfloor na \rfloor} \xrightarrow{n \rightarrow \infty} e^{-\lambda n},$$

which means that $\frac{1}{n} \text{Geo}(\lambda n^{-1}) \xrightarrow{d} \text{Exp}(\lambda)$.

Example 12.4. Define

$$F_n(a) = \begin{cases} \left(\frac{a}{1+a}\right)^n & a > 0, \\ 0 & a \leq 0. \end{cases}$$

For all $a \in \mathbb{R}$ we have that $F_n(a) \xrightarrow{n \rightarrow \infty} 0$, and since 0 is not a distribution function for any random variable X , the sequence F_n does not converge in distribution.

12.2 The law of large numbers

The law of large numbers is describing the intuitive phenomenon that the more times a certain experiment is repeated, the more the result of the experiment converges to its expectation.

Definition 12.7 (i.i.d. sequence). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X_n(n \in \mathbb{N}): \Omega \rightarrow \mathbb{R}$ be a sequence of random variables. We say that $\{X_n\}_{n=1}^{\infty}$ is an i.i.d. (independent and identically distributed) sequence if they are independent and there exists a distribution function F so that $F_{X_n} = F$ for all $n \geq 1$.

Remark 12.6. Since expectation, variance, and functions of random variables depend only on the random variable's distribution, we get that these are all identical for any random variable in an i.i.d. sequence of random variables.

Theorem 12.1 (The weak law of large numbers). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X_n(n \in \mathbb{N}): \Omega \rightarrow \mathbb{R}$ be an i.i.d. sequence of random variables with finite expectation $\mu \in \mathbb{R}$. Then weak law of large numbers (also called Khinchin's law) states that*

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mu.$$

Theorem 12.2 (The strong law of large numbers). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X_n(n \in \mathbb{N}): \Omega \rightarrow \mathbb{R}$ be an i.i.d. sequence of random variables with finite expectation $\mu \in \mathbb{R}$. Then strong law of large numbers states that*

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{a.s.} \mu.$$

Remark 12.7. The weak and strong laws of large numbers have not very easy proofs, and thus we will only provide a weaker version of the weak law of large numbers. There is also a proof for a weaker version of the strong law of large numbers in the notes, where we assume that the variables have finite fourth moments. This really emphasizes how the existence of higher moments helps us with probabilistic computations.

Theorem 12.3 (The weak law of large numbers of variables with second moments). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X_n (n \in \mathbb{N}) : \Omega \rightarrow \mathbb{R}$ be an i.i.d. sequence of random variables with finite expectation μ , and finite variance σ^2 . Then*

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\mathcal{L}^2} \mu \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mu \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{d} \mu.$$

Remark 12.8. As we saw in Remark 12.3, convergence in $\mathcal{L}^2$ implies convergence in probability, which in turn implies convergence in distribution. Before the proof of Theorem 12.3, we will quickly give a short prove of the first implication. The second implication is more difficult to prove, and its proof will unfortunately be omitted.

Proposition 12.4. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X, X_n (n \in \mathbb{N}) : \Omega \rightarrow \mathbb{R}$ be random variables such that $X_n \xrightarrow{\mathcal{L}^2} X$. Then $X_n \xrightarrow{p} X$.*

Proof. We know that $\lim_{n \rightarrow \infty} \mathbf{E}[(X_n - X)^2] = 0$, so by applying Markov's inequality, for all $\epsilon > 0$ we get that

$$\mathbf{P}(|X_n - X| \geq \epsilon) = \mathbf{P}(|X_n - X|^2 \geq \epsilon^2) \leq \frac{\mathbf{E}[|X_n - X|^2]}{\epsilon^2} \xrightarrow{n \rightarrow \infty} 0,$$

which completes the proof. □

Proof of Theorem 12.3. We will only prove convergence in $\mathcal{L}^2$. From linearity of expectation we have that

$$\mathbf{E} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbf{E}[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu.$$

Since $\{X_n\}_{n=1}^\infty$ are independent, from properties of variance we get that

$$\text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}.$$

Therefore,

$$\begin{aligned} \mathbf{E} \left[\left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right)^2 \right] &= \mathbf{E} \left[\left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbf{E} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] \right)^2 \right] \\ &= \text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{\sigma^2}{n} \xrightarrow{n \rightarrow \infty} 0, \end{aligned}$$

which proves convergence in $\mathcal{L}^2$, and completes the proof. □

12.3 The law of truly large numbers

Informally, the law of truly large numbers states that in an infinite sequence of independent experiments, if an event has a positive probability (small as it may be), then it will be observed infinitely many times throughout the sequence. For example, if we flipped a fair coin an infinite amount of times in a row, it is natural to ask what is the probability that at some point we will observe 10^{100} heads in a row. Since the probability of this event is at least $2^{-10^{100}} > 0$, the law of truly large numbers says that the probability of seeing 10^{100} heads in a row is 1. We will prove something similar for a different example.

Example 12.5. Suppose we are flipping a fair coin an infinite amount of times. We want to show that we will eventually get heads 10^{100} times in a row. Instead of showing this, we will show that we will eventually get heads 4 times in a row (the proof for 10^{100} is similar). Define

$$A_n := \{\text{results } n, n+1, n+2, n+3 \text{ are heads}\}.$$

Let $X_n := \mathbb{1}_{A_n}$. Note $A_1, A_5, A_9, \dots$ are independent, and denote $B_n = X_{4(n-1)+1}$. We have that $\mathbf{P}(B_n) = 2^{-4} > 0$, and therefore $\sum_{k=1}^{\infty} \mathbf{P}(B_k) = \infty$. Since the events $\{B_n\}_{n=1}^{\infty}$ are independent, by the second Borel–Cantelli lemma we get that $\mathbf{P}(\limsup_{n \rightarrow \infty} B_n) = 1$. This shows that the probability of eventually getting heads 4 times in a row is 1.

Remark 12.9. Recall that for any natural number $b \geq 2$, each number $x \in \mathbb{R}$ has a representation in base b , and that this representation is unique, up to a countable set of numbers.

Definition 12.8 (Normal number). Let $b \geq 2$ be an integer. A real number $x \in [0, 1]$ is said to be normal in base b if for every integer $s \geq 1$ and for every block $(d_1, d_2, \dots, d_s) \in \{0, 1, \dots, b-1\}^s$, the frequency with which this block appears in the base- b expansion of x exists and equals b^{-s} . Equivalently, writing $x = 0.d_1d_2d_3\dots$ in base b , we require that

$$\lim_{N \rightarrow \infty} \frac{\left| \{1 \leq n \leq N \mid (d_n, d_{n+1}, \dots, d_{n+s-1}) = (a_1, \dots, a_s)\} \right|}{N} = b^{-s},$$

for every choice of digits $(a_1, \dots, a_s) \in \{0, 1, \dots, b-1\}^s$.

Proposition 12.5. Let $\Omega = [0, 1]$ and let $\mathbf{P}$ be the uniform probability function on Ω . Then for all $b \geq 2$ we have

$$\mathbf{P}\left(\{\omega \in \Omega \mid \omega \text{ is normal in base } b\}\right) = 1.$$

In particular, since the number of possible bases is countable, we get that

$$\mathbf{P}\left(\{\omega \in \Omega \mid \omega \text{ is normal in base } b \text{ for every natural } b \geq 2\}\right) = 1.$$

This implies that almost every $x \in [0, 1]$ is normal in every base.

Proof. We will only show the proof for the case $s = 1$. That is, we will show that for any $b \geq 2$ and $s \in \{0, 1, \dots, b-1\}$ we have

$$\lim_{N \rightarrow \infty} \frac{\left| \{1 \leq n \leq N \mid d_n = s\} \right|}{N} = \frac{1}{b},$$

in probability 1 (almost all $\omega \in \Omega$). Indeed, for any $b \geq 2$ we define $X_n^{b,\omega} : \Omega \rightarrow \mathbb{R}$ by $X_n^{b,\omega}(\omega) = \mathbb{1}_{d_n^{b,\omega} = s}$. It is clear that $\{X_n^{b,\omega}\}$ is an i.i.d. sequence, and therefore, by the strong law of large numbers,

$$\lim_{N \rightarrow \infty} \frac{\left| \{1 \leq n \leq N \mid d_n = s\} \right|}{N} = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N X_k^{b,\omega} = \mathbf{E}\left[X_1^{b,\omega}\right] = \frac{1}{b},$$

in probability 1, which completes the proof. $\square$

Remark 12.10. Although we can show that almost every number is normal in every base, it is very difficult to show that a given number is normal. For example we do not know whether $\sqrt{2}$, π , or e are normal. We do know that all rational numbers are not normal (and the proof is immediate), and we have constructed certain normal numbers, but there isn't any easy way to tell whether a number is normal or not.

13 The central limit theorem

13.1 Introduction to the theorem

Theorem 13.1 (Central limit theorem). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X_n: \Omega \rightarrow \mathbb{R}$ be a sequence of i.i.d. random variables with expectation $\mathbf{E}[X_1] = \mu < \infty$, and variance $\text{Var}(X_1) = \sigma^2 < \infty$. Then*

$$\frac{1}{\sqrt{n\sigma^2}} \sum_{k=1}^n (X_k - \mu) \xrightarrow{d} N(0, 1).$$

Remark 13.1. For a sufficiently large $n \geq 1$, the law of large numbers tells us that for an i.i.d. sequence of random variables $\{X_n\}_{n=1}^\infty$, we have the approximation $\sum_{k=1}^n X_k \approx n\mu$. The central limit theorem tells us that if the random variables also have finite expectation μ and finite variance σ^2 , then

$$\sum_{k=1}^n X_k \approx n\mu + \sqrt{n\sigma^2}Y,$$

for $Y \sim N(0, 1)$. Another way to view the approximation is in the form

$$\frac{1}{n} \sum_{k=1}^n X_k \approx Y,$$

for $Y \sim N(\mu, \sigma^2/n)$. The general consensus is that any $n \geq 30$ is a sufficiently large n , in the sense that the difference between these approximations and the real value is insignificant.

Remark 13.2. Sometimes, we may have a discrete variable $M = \frac{1}{N} \sum_{j=1}^N X_j$, so that $M \sim N(0, 1)$, and we will be asked to compute $\mathbf{P}(M = 0)$. Since M is discrete we possibly have $\mathbf{P}(M = 0) > 0$, however $N(0, 1)$ is continuous and so by that approximation we will always get $\mathbf{P}(M = 0) = 0$, which is bad. The best way to get around this, is to recall that $N(0, 1)$ is a continuous approximation of M , and therefore, if for example M can get values in $\mathbb{Z}$, we would have $\mathbf{P}(M = 0) = \mathbf{P}(-0.5 \leq M \leq 0.5) \approx 2\Phi(0.5) - 1$. The reason we choose $M \in [-0.5, 0.5]$, and not some other interval, is because M may get values in -1 or 1 , so the best approximation would be to take the interval which includes the points that are closer to 0 than any other value in $\mathbb{Z}$. If we knew that M only gets values in $2\mathbb{Z}$, computing $\mathbf{P}(-1 \leq M \leq 1)$ would be a better approximation. An example is given in Exercise 13.4.

Exercise 13.1. Suppose that the expected score of all students in a course is $\mu = 80$, and that the standard deviation is $\sigma = 20$. Find an approximation for the average score of 100 students assuming their scores are independent. Then find an approximation for the probability that the average is less than 70.

Solution. Denote by X_k the score of the k th student. The average score in the course is then

$$M := \frac{1}{100} \sum_{k=1}^{100} X_k.$$

Since 100 is a sufficiently large number (certainly, it's greater than 30), we get that the distribution of the sample mean is $M \sim N(80, 4)$, and that the normalized distribution of the sample mean distributes $M_0 := \frac{1}{100} \sum_{k=1}^{100} (X_k - 80) \sim N(0, 4)$. If we denote $M_N := M_0 / \sqrt{\frac{\sigma^2}{n}} = M_0/2$ we get that $M_N \sim N(0, 1)$. Finally, we can approximate the probability that the average is less than 70 by definition:

$$\mathbf{P}(M < 70) = \mathbf{P}(M_0 < -10) = \mathbf{P}(M_N < -5) = \Phi(5) \leq e^{-12},$$

where the last inequality was done by considering a standard table of the approximate values of the cumulative distribution function of standard deviation. This tells us that the probability of the average score in class to be under 70 is essentially zero (but definitely more than zero).

Exercise 13.2. Let $\{X_k\}_{k=1}^{\infty}$ be a sequence of i.i.d. random variables, such that $\mathbf{E}[X_1] = \mu$ and $\text{Var}(X_1) = 1$. Show that there exist $N_0 \in \mathbb{N}$ and $c_0 > 0$ so that for all $n \geq N_0$ we have

$$\mathbf{P} \left(\left| \sum_{k=1}^n (X_k - \mu) \right| \leq \frac{1}{2} \sqrt{n} \right) \geq c_0.$$

Solution. We notice that for large n we have $\sum_{k=1}^n (X_k - \mu) \sim N(0, n)$, and thus $\frac{1}{\sqrt{n}} \sum_{k=1}^n (X_k - \mu) \sim N(0, 1)$. From the central limit theorem for sufficiently large n we get

$$\mathbf{P} \left(\left| \sum_{k=1}^n (X_k - \mu) \right| \leq \frac{1}{2} \sqrt{n} \right) = \mathbf{P} \left(\left| \frac{1}{\sqrt{n}} \sum_{k=1}^n (X_k - \mu) \right| \leq \frac{1}{2} \right) \approx 2\Phi(0.5) - 1 > 0,$$

so for sufficiently large n the value $c_0 = (2\Phi(0.5) - 1)/2$ works.

Exercise 13.3. Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. random variables, so that $\mathbf{E}[X_1] = 60$ and $\text{Var}(X_1) = 25$. Show that there exists a sufficiently large N so that the probability that $\frac{1}{N} \sum_{k=1}^N X_k \in [55, 65]$ is greater than 0.99.

Solution. Denote $\mu = 60$ and $\sigma^2 = 25$. Note that

$$\mathbf{P} \left(55 \leq \frac{1}{N} \sum_{k=1}^N X_k \leq 65 \right) = \mathbf{P} \left(-5 \leq \frac{1}{N} \sum_{k=1}^N (X_k - 60) \leq 5 \right).$$

We know that for sufficiently large N that $M_0 := \frac{1}{N} \sum_{k=1}^N (X_k - 60) \sim N(0, \frac{25}{N})$, and therefore $M_N := \frac{1}{5\sqrt{N}} \sum_{k=1}^N (X_k - 60) \sim N(0, 1)$. We have that

$$\mathbf{P} \left(-5 \leq \frac{1}{N} \sum_{k=1}^N (X_k - 60) \leq 5 \right) = \mathbf{P} \left(-\sqrt{N} \leq \frac{1}{5\sqrt{N}} \sum_{k=1}^N (X_k - 60) \leq \sqrt{N} \right).$$

Now since we want to bound this expression from below we notice that for any $M > 0$ so that $N \geq M^2$ we have

$$\mathbf{P} \left(-\sqrt{N} \leq \frac{1}{5\sqrt{N}} \sum_{k=1}^N (X_k - 60) \leq \sqrt{N} \right) \geq \mathbf{P} \left(-M \leq \frac{1}{5\sqrt{N}} \sum_{k=1}^N (X_k - 60) \leq M \right),$$

and that by the symmetry of $M_N \sim N(0, 1)$ around $x = 0$ we get that the right hand side converges to $2\Phi(M) - 1$ as $N \rightarrow \infty$. Now, notice that we can't immediately take the limit on both sides, because the left hand side does not necessarily converge, so take $\liminf$ on both sides and get

$$\liminf_{N \rightarrow \infty} \mathbf{P} \left(-\sqrt{N} \leq \frac{1}{5\sqrt{N}} \sum_{k=1}^N (X_k - 60) \leq \sqrt{N} \right) \geq 2\Phi(M) - 1.$$

By taking the limit $M \rightarrow \infty$ on both sides we get that

$$\liminf_{N \rightarrow \infty} \mathbf{P} \left(-\sqrt{N} \leq \frac{1}{5\sqrt{N}} \sum_{k=1}^N (X_k - 60) \leq \sqrt{N} \right) \geq 1,$$

and therefore

$$\lim_{N \rightarrow \infty} \mathbf{P} \left(-\sqrt{N} \leq \frac{1}{5\sqrt{N}} \sum_{k=1}^N (X_k - 60) \leq \sqrt{N} \right) = 1,$$

which completes the proof.

Exercise 13.4. Suppose that there is a game in a casino where each game you win $\text{Geo}(0.1)$ coins, and lose $\text{Poi}(10)$ coins, where the wins and the losses are independent. A player plays that game repeatedly, where the sequence of games is independent. What is the probability that the player gained more than 20 coins after 100 games (counting losses)? What is the probability that his wins and losses exactly cancel out?

Solution. Denote by X_k and Y_k the amount of coins he won and lost in the k th game respectively. We want to compute

$$S = \sum_{j=1}^{100} (X_j - Y_j),$$

where $X_j \sim \text{Geo}(0.1)$ and $Y_j \sim \text{Poi}(10)$, and all random variables $X_1, \dots, X_{100}, Y_1, \dots, Y_{100}$ are independent. Note that from linearity

$$\mathbf{E}[X_j - Y_j] = \mathbf{E}[X_j] - \mathbf{E}[Y_j] = 10 - 10 = 0,$$

so denote $\mu = 0$, and from properties of variance we have

$$\text{Var}(X_j - Y_j) = \text{Var}(X_j) + \text{Var}(Y_j) - \underbrace{2 \text{Cov}(X_j, Y_j)}_{=0} = \frac{0.9}{(0.1)^2} + 10 = 100,$$

so denote $\sigma^2 = 100$. Since 100 is sufficiently large ($100 > 30$) the central limit theorem implies that $S \sim N(0, 10000)$ and thus $S/100 \sim N(0, 1)$. This means that

$$\mathbf{P}(S > 20) \approx \mathbf{P}(S/100 > 0.2) = 1 - \Phi(0.2) \approx 0.4207.$$

Now we are being ask to calculate $\mathbf{P}(S = 0)$. Since S is discrete and get values in $\mathbb{Z}$, we can approximate by

$$\mathbf{P}(S = 0) = \mathbf{P}(-0.5 \leq S \leq 0.5) = \mathbf{P}(-0.005 \leq S/10 \leq 0.005) = 2\Phi(0.005) - 1 \approx 0.00398.$$

13.2 Preliminaries for the proof

In this subsection, the goal is to introduce the tools required for the proof of the central limit theorem. We will not provide almost any proofs, but the definitions and theorems are worth reading in case you are interested in the field, or in the proof of the central limit theorem.

Definition 13.1 (Characteristic function). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let X be a random variable. The characteristic function of X , denoted $\varphi_X: \mathbb{R} \rightarrow \mathbb{C}$ is defined by

$$\varphi_X(t) := \mathbf{E} \left[e^{itX} \right] = \mathbf{E} [\cos(tX) + i \sin(tX)] = \mathbf{E} [\cos(tX)] + i \mathbf{E} [\sin(tX)].$$

Remark 13.3. The function $\varphi_X(t)$ reminds the moment generating function $M_X(t) = \mathbf{E}[e^{tX}]$. Indeed, $\varphi_X(t)$ is the Fourier transform of the function X , and $M_X(t)$ is the Laplace transform of X . While the the characteristic function function $\varphi_X(t)$ is well defined for all $t \in \mathbb{R}$, the moment generating function $M_X(t)$ is not defined for all random variables, and even when it is defined it may not be defined for all $t \in \mathbb{R}$.

Remark 13.4. The reason that $\varphi_X(t) = \mathbf{E}[e^{itX}]$ is always defined is because the function $x \mapsto e^{itx}$ is a continuous and bounded function, so the expectation of e^{itX} is always finite.

Remark 13.5. We have the following correspondences between characteristic functions and distributions, which we can derive straight from the definition, with a bit (or slightly more than a bit) computation:

$X \sim$	$\text{Bin}(n, p)$	$\text{Geo}(p)$	$\text{Poi}(\lambda)$	$U([a, b])$	$\text{Exp}(\lambda)$	$N(0, 1)$
$\varphi_X(t)$	$(1 - p + pe^{it})^n$	$\frac{pe^{it}}{1 - (1-p)e^{it}}$	$\exp(\lambda(e^{it} - 1))$	$\frac{e^{itb} - e^{ita}}{it(b-a)}$	$\frac{\lambda}{\lambda - it}$	$\exp(-t^2/2)$

Proposition 13.2 (Properties of the characteristic function). *Let $X, Y: \Omega \rightarrow \mathbb{R}$ be random variables, and let $a, b \in \mathbb{R}$ be real numbers. Then*

- (1) $\varphi_X(0) = 1$.
- (2) $t \mapsto \varphi_X(t)$ is a uniformly continuous function.
- (3) $\varphi_{aX+b}(t) = e^{ib}\varphi_X(at)$ for all $t \in \mathbb{R}$.
- (4) $\varphi_X(t) = \overline{\varphi_X(-t)}$ for all $t \in \mathbb{R}$.
- (5) if $\mathbf{E}[|X|^k] < \infty$ for $k \in \mathbb{N}$, then $\varphi_X(t)$ is differentiable k times and $\varphi_X^{(m)}(0) = i^m \mathbf{E}[X^m]$ for all $0 \leq m \leq k$.
- (6) If X and Y are independent, then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$ for all $t \in \mathbb{R}$.

Proposition 13.3. *Let $X, Y: \Omega \rightarrow \mathbb{R}$ be random variables, Then $\varphi_X = \varphi_Y$ if and only if $F_X = F_Y$.*

Theorem 13.4 (Lévy's continuity theorem). *Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables (not necessarily sharing a probability space), and let X be another random variable. Then $X_n \xrightarrow{d} X$ if and only if $\lim_{n \rightarrow \infty} \varphi_{X_n}(t) = \varphi_X(t)$ for all $t \in \mathbb{R}$.*

Proposition 13.5. *Let X be a random variable with $\mathbf{E}[X] = 0$ and $\mathbf{E}[X^2] = 1$. Then,*

$$\varphi_X(t) = 1 - \frac{t^2}{2} + o(t^2),$$

when $t \rightarrow 0$.

13.3 Miscellaneous remarks and definitions

Definition 13.2 (Standard score). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X_1, \dots, X_N$ be i.i.d. random variables. Then the sample mean of $X_1, \dots, X_N$ is defined by

$$\bar{X} = \frac{1}{N} \sum_{k=1}^N X_k.$$

Intuitively, we can think of each X_k as a sample from an experiment, and then $\bar{X}$ is the mean of the samples. The weak and strong laws of large numbers imply that it converges to the mean (in probability and almost surely respectively).

Definition 13.3 (Standard score). Let $X: \Omega \rightarrow \mathbb{R}$ be a random variable with finite expectation and finite variance. Then we define the standard score (or z-score) of some $x \in \mathbb{R}$ by

$$z(x) = \frac{x - \mathbf{E}[X]}{\sqrt{\text{Var}(X)}} = \frac{x - \mathbf{E}[X]}{\sigma_X},$$

where σ_X is the standard deviation of X .

Remark 13.6 (68–95–99.7 rule). If $X \sim N(\mu, \sigma)$, then $\mathbf{P}(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 68.27\%$ and $\mathbf{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.45\%$, and $\mathbf{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.73\%$. This implies that the standard scores of samples is almost always in the range $[-3, 3]$.

Definition 13.4 (T-score). Given a sample x , we define the t-score (or *hensachi*) of x by normalizing its z -score as follows: $t(z) = 50 + 10z(x)$.

Definition 13.5 (Discrete Markov chain). Let $X_0, X_1, \dots$ be random variables with the *Markov property*. That is,

$$\mathbf{P}(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \mathbf{P}(X_{n+1} = x \mid X_n = x_n),$$

for all values such that the equality is well defined, that is, $\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) > 0$. Then the sequence $X_0, X_1, \dots$ is called a discrete Markov chain. The possible values of X_i form a countable set S called the *state space* of the chain.

Remark 13.7 (Discrete Markov chain intuition). Intuitively, a Markov chain describes processes that change over time. For example, if we denote the weather of each day starting from today by $X_0, X_1, \dots$, and assume the weather of the k th day only depends on the weather of the $(k - 1)$ th day, that would be a discrete Markov chain.

14 Random walks

In this section we will generalize the theory of random walks (which we saw in Section 3.3).

14.1 Simple symmetric random walks

A random walk is called simple if the size of every step is exactly one unit, and it is called symmetric if the probability of going in any one direction is equal to the probability of going in the exact opposite direction. If a random walk is simple and symmetric we sometimes call it SSRW instead.

Definition 14.1 (One-dimensional SSRW in $\mathbb{Z}$). Let $\{X_n\}_{n=1}^\infty$ be i.i.d. random variables with $\mathbf{P}(X_n = 1) = \mathbf{P}(X_n = -1) = \frac{1}{2}$. Define $S_n = \sum_{i=1}^n X_i$ for all $n \geq 1$, and set $S_0 = 0$. Then the sequence $\{S_n\}_{n=0}^\infty$ is called a one-dimensional simple symmetric random walk.

Definition 14.2 (d -dimensional SSRW in $\mathbb{R}^d$). Suppose $\{X_n^{(d)}\}_{n=1}^\infty$ are i.i.d. random vectors with $\mathbf{P}(X_n^{(d)} = e_j) = \mathbf{P}(X_n^{(d)} = -e_j) = \frac{1}{2d}$ for all $1 \leq j \leq d$ (where e_j is the unit vector in the j th dimension). Let $S_n^{(d)} = \sum_{i=1}^n X_i^{(d)}$ for all $n \geq 1$, and set $S_0^{(d)} = 0$. Then the sequence $\{S_n^{(d)}\}_{n=0}^\infty$ is called a d -dimensional simple symmetric random walk.

Definition 14.3 (Recurrence). For a SSRW $\{S_n^{(d)}\}_{n=0}^\infty$ set $R_d = \{S_n^{(d)} = 0 \text{ for some } j\} = \{\text{The SSRW returns to the origin}\}$. We define $p_d = \mathbf{P}(R_d)$.

Definition 14.4 (Recurrent and transient SSRW). Let $\{S_n^{(d)}\}_{n=0}^\infty$ be a SSRW. If $p_d = 1$ we say that it is *recurrent*. If $p_d < 1$ we say that it is *transient*.

Exercise 14.1. For what $d \geq 1$ the SSRW $\{S_n^{(d)}\}_{n=0}^\infty$ is recurrent?

Solution. This is a slightly too difficult exercise, we have to start by finding a simpler characterization for the condition $p_d < 1$. Let:

$$N_d = \sum_{n=1}^{\infty} \mathbb{1}_{S_n^{(d)}=0} = \# \text{Number of visits to the origin.}$$

By linearity of expectation:

$$\mathbf{E}[N_d] = \mathbf{E} \left[\sum_{n=1}^{\infty} \mathbb{1}_{S_n^{(d)}=0} \right] = \sum_{n=1}^{\infty} \mathbf{E}[\mathbb{1}_{S_n^{(d)}=0}] = \sum_{n=1}^{\infty} \mathbf{P}(S_n^{(d)} = 0) = \sum_{n=1}^{\infty} \mathbf{P}(S_{2n}^{(d)} = 0),$$

where the last transition follows from the fact any visit to the origin must be on an even step.

Next, assume that $p_d < 1$. Since the steps are independent, and we can look at each visit to the origin as a “reset”, we can easily see that $\mathbf{P}(N_d = 0) = 1 - p_d$ and $\mathbf{P}(N_d = 1) = p_d(1 - p_d)$, and in general $\mathbf{P}(N_d = m) = p_d^m(1 - p_d)$. This implies that $N_d \sim \text{Geo}(1 - p_d)$ with values in $0, 1, 2, \dots$. In particular, $\mathbf{E}[N_d] < \infty$.

On the other hand, assume that $p_d = 1$. In this case $\mathbf{P}(N_d = \infty) = 1$ hence $\mathbf{E}[N_d] = \infty$. This implies that $\mathbf{E}[N_d] < \infty$ implies $p_d < 1$.

Concluding the last two paragraphs we get that:

$$p_d < 1 \iff \sum_{n=1}^{\infty} \mathbf{P}(S_{2n}^{(d)} = 0) < 1.$$

Proposition 14.1. One and two dimensional SSRW are recurrent.

Proof. First let $d = 1$. Using Stirling's approximation, given by $n! \sim n^n e^{-n} \sqrt{2\pi n}$, we have that:

$$\mathbf{P}\left(S_{2n}^{(1)} = 0\right) = \frac{\binom{2n}{n}}{2^{2n}} = \frac{(2n)!}{(n!)^2 2^{2n}} \sim \frac{(2n)^{2n} e^{-2n} \sqrt{2\pi(2n)}}{2^{2n} n^{2n} e^{-2n} (2\pi n)} = \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{n}}.$$

So since $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} = \infty$, we have that $\sum_{n=1}^{\infty} \mathbf{P}\left(S_{2n}^{(d)} = 0\right) = \infty$, hence by the result in Exercise 14.1 we conclude that $p_1 = 1$.

Next let $d = 2$. We have that:

$$\begin{aligned} \mathbf{P}\left(S_{2n}^{(2)} = 0\right) &= \frac{\sum_{k=0}^n \binom{2n}{k} \binom{2n-k}{n-k} \binom{2n-2k}{n-k}}{4^{2n}} \\ &= \frac{1}{4^{2n}} \sum_{k=0}^n \frac{(2n)!}{k! k! (n-k)! (n-k)!} \\ &= \frac{1}{4^{2n}} \frac{(2n)!}{(n!)^2} \sum_{k=0}^n \left(\frac{n!}{k!(n-k)!}\right)^2 \\ &= \frac{1}{4^{2n}} \frac{(2n)!}{(n!)^2} \sum_{k=0}^n \binom{n}{k}^2 \\ &= \frac{\binom{2n}{n}^2}{4^{2n}} = \left(\frac{\binom{2n}{n}}{2^{2n}}\right)^2. \end{aligned}$$

Thus, by the same Stirling approximation result from the one-dimensional case we get:

$$\mathbf{P}\left(S_{2n}^{(2)} = 0\right) = \left(\frac{\binom{2n}{n}}{2^{2n}}\right)^2 = \mathbf{P}\left(S_{2n}^{(1)} = 0\right)^2 \sim \left(\frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{n}}\right)^2 = \frac{1}{\pi n}.$$

Hence $\sum_{n=1}^{\infty} \mathbf{P}\left(S_{2n}^{(2)} = 0\right) = \infty$, and so by Exercise 14.1 we conclude that $p_2 = 1$, which completes the proof. $\square$

Remark 14.1. A three dimensional random walk is transient. It is not the case that the proof is omitted because we do not have the tools to show it yet. It is just that the proof is purely algebraic and really technical. Eventually we get a bound of the form $\mathbf{P}(S_{2n}^{(3)}) \leq C \cdot n^{-3/2}$ for some $C > 0$, hence $\sum_{n=1}^{\infty} \mathbf{P}\left(S_{2n}^{(2)} = 0\right) < \infty$, and so $p_d < 1$.

14.2 One-dimensional SSRW

In this subsection, we only consider one-dimensional SSRW, but we generalize them by setting $S_0 = j$ (this effectively means that the walk starts from j instead of the origin). In the case of a SSRW starting at j we denote the probability function by P_j and the expectation by E_j (concretely $\mathbf{E}_j[Y] = \mathbf{E}[Y \mid S_0 = j]$). We also declare $\tau_k := \min\{n \geq 0 \mid S_n = k\}$, that is, τ_j is the first step we visit j .

Proposition 14.2. *We have that $\mathbf{P}_j(\tau_0 < \tau_M) = \frac{M-j}{M}$ and $\mathbf{P}_j(\tau_M < \tau_0) = \frac{j}{M}$ for all $0 \leq j \leq n$.*

Using first step analysis. Let $h_j = \mathbf{P}_j(\tau_0 < \tau_M)$ so that $h_0 = 1$ and $h_M = 0$. Let us denote by R the event that the first step was “going right”, and by L the event that the first step was “going left”. Then by the law of total probability we have:

$$h_j = \frac{1}{2} \mathbf{P}_j(\tau_0 < \tau_M \mid L) + \frac{1}{2} \mathbf{P}_j(\tau_0 < \tau_M \mid R) = \frac{1}{2} h_{j-1} + \frac{1}{2} h_{j+1},$$

which implies that $h_{j-1} - h_j = h_j - h_{j+1}$ for all $1 \leq j \leq M-1$ (this process is called “first step analysis”). This implies that $\{h_j\}$ is an arithmetic progression with step $\gamma = h_{j-1} - h_j$. Thus, it follows that:

$$1 - 0 = h_0 - h_M = \sum_{j=1}^M (h_{j-1} - h_j) = \gamma M,$$

so $\gamma = M^{-1}$, so we conclude that:

$$h_j = \sum_{k=j}^{M-1} (h_k - h_{k+1}) = (M-j)\gamma = \frac{M-j}{M},$$

which completes the first part of the proof, but the second part is immediate so the proof is complete. $\square$

Proposition 14.3. *Let $\tau_{0,M} = \min \{n \geq 0 \mid S_n = 0 \text{ or } S_n = M\}$. Then $E_j[\tau_{0,M}] = j(M-j)$ for $0 \leq j \leq M$.*

Proof. Let $h_j = \mathbf{E}_j[\tau_{0,M}]$. Then notice that $h_0 = h_M = 0$. From first step analysis we get the relation:

$$h_{j+1} - h_j = h_j - h_{j-1} - 2,$$

and by setting $\gamma = h_1 - h_0 = h_1$ we get that $h_2 - h_1 = \gamma - 2$ and $h_3 - h_2 = h_2 - h_1 - 2 = \gamma - 4$, and similarly:

$$h_{j+1} - h_j = \gamma - 2j,$$

for $1 \leq j \leq M-1$. Since:

$$0 = h_M - h_0 = \sum_{j=0}^{M-1} (h_{j+1} - h_j) = \sum_{j=0}^{M-1} (\gamma - 2j) = \gamma M - 2 \frac{M(M-1)}{2},$$

we conclude that $\gamma = M-1$, so from the previous relation we get that:

$$h_{j+1} - h_j = \gamma - 2j = M-1-2j,$$

for $1 \leq j \leq M-1$. Using this relation, we get that:

$$h_j = \sum_{k=0}^{j-1} (h_{k+1} - h_k) = \sum_{k=0}^{j-1} (M-1-2k) = (M-1)j + \frac{2(j-1)j}{2} = j(M-1-(j-1)) = j(M-j),$$

which concludes the proof. $\square$

Remark 14.2. We define $\hat{\tau}_0 = \inf \{n \geq 1 \mid S_n = 0\}$. By Theorem 14.1 we know that $\mathbf{P}(\hat{\tau}_0 < \infty) = 1$. Note that since $\mathbf{E}_0[\hat{\tau}_0] \geq \mathbf{E}_1[\tau_{0,M}] = M-1$ for any M , we get that $\mathbf{E}_0[\hat{\tau}_0] = \infty$.

14.3 Non-symmetric simple random walk

In this subsection we consider a simple random walk on $\mathbb{Z}$, that is not symmetric. That is, $\mathbf{P}(X_n = 1) = p$ and $\mathbf{P}(X_n = -1) = 1-p$ for $p \in (0, 1) \setminus \left\{\frac{1}{2}\right\}$. We use the same notation from the previous subsection for walks that start from j .

Proposition 14.4. *Let $\frac{1}{2} < p < 1$. Then we have the identity:*

$$\mathbf{P}_j(\tau_0 < \tau_M) = \frac{\left(\frac{1-p}{p}\right)^j - \left(\frac{1-p}{p}\right)^M}{1 - \left(\frac{1-p}{p}\right)^M}.$$

Corollary 14.5. $\mathbf{P}_j(\tau_0 < \infty) = \left(\frac{1-p}{p}\right)^j$.

Proof. Since $\{\tau_0 < \infty\} = \bigcup_{m=1}^{\infty} \{\tau_0 < \tau_M\}$ we have that:

$$\mathbf{P}_j(\tau_0 < \infty) = \lim_{m \rightarrow \infty} \mathbf{P}_j(\tau_0 < \tau_M) = \lim_{m \rightarrow \infty} \frac{\left(\frac{1-p}{p}\right)^j - \left(\frac{1-p}{p}\right)^M}{1 - \left(\frac{1-p}{p}\right)^M} = \left(\frac{1-p}{p}\right)^j.$$

□

Proof of Theorem 14.4. Let $h_j = \mathbf{P}_j(\tau_0 < \tau_M)$ for $0 \leq j \leq M$, so that $h_0 = 1$ and $h_M = 0$. Using first step analysis we get the relation:

$$p(h_j - h_{j+1}) = (1-p)(h_{j-1} - h_j).$$

If we set $\gamma = h_0 - h_1$ then we can see that in general this gives the relation:

$$h_j - h_{j+1} = \left(\frac{1-p}{p}\right)^j \gamma,$$

for $1 \leq j \leq M-1$. Using this relation, we get that:

$$\begin{aligned} h_j - h_m &= \sum_{k=j}^{m-1} (h_k - h_{k+1}) = \sum_{k=j}^{m-1} \left(\frac{1-p}{p}\right)^k \gamma \\ &= \gamma \frac{\left(\frac{1-p}{p}\right)^j - \left(\frac{1-p}{p}\right)^m}{1 - \frac{1-p}{p}} = \frac{\gamma p}{2(p-1)} \left(\left(\frac{1-p}{p}\right)^j - \left(\frac{1-p}{p}\right)^m \right). \end{aligned}$$

For $j = 0$ we get

$$1 = \frac{\gamma p}{2(p-1)} \left(1 - \left(\frac{1-p}{p}\right)^m \right),$$

and thus:

$$\gamma = \frac{2(p-1)}{p} \left(1 - \left(\frac{1-p}{p}\right)^m \right)^{-1}.$$

Now by substituting γ in the formula for h_j we saw earlier we get:

$$\begin{aligned} h_j &= \frac{\gamma p}{2(p-1)} \left(\left(\frac{1-p}{p}\right)^j - \left(\frac{1-p}{p}\right)^m \right) \\ &= \frac{\left(\frac{1-p}{p}\right)^j - \left(\frac{1-p}{p}\right)^m}{1 - \left(\frac{1-p}{p}\right)^m}. \end{aligned}$$

This completes the proof. □