Introduction to Probability Theory

1 Probability Spaces

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 seperately.

First, let us introduce a paradox.

Paradox 1.1. (Bertrand's Paradox). *Consider an equilateral triangle inscribed in 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?*

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.

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

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 $\Omega = \{1,2,\ldots,11,12\}$ 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 $\mathbf{P}\colon \Omega \to [0,1]$ 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. A subset of the sample space $A \subseteq \Omega$ is called an event. We also define:

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

Here are a few properties of 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
$$\mathbf{P}\left(\cup_{n=1}^n A_n\right) = \sum_{n=1}^N \mathbf{P}(A_n).$$
5. If $\{A_n\}_{n=1}^\infty$ is a sequence of pairwise disjoint sets then
$$\mathbf{P}\left(\cup_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty \mathbf{P}(A_n).$$

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

We now proceed to consider an experiment in which we choose a direction in $\mathbb{R}^2$ at random, on S^1 and write it. The sample space is:

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

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 $\mathbf{P}([a, b]) = b - a$. The answer is that with the definition we have worked with so far, we can't. We see that $\mathbf{P}(\{a\}) = 0$ for any $a \in S^1$, and thus we have that

$$\mathbf{P}(\Omega) = \sum_{\omega \in \Omega} \mathbf{P}(\omega) = 0.$$

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

Lemma 1.1. *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 $\mathbf{P}(E) = a$ then we get

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

and this equation has no solution.

The classical solution to this problem, is to only define the probability function only on certain subsets of the sample space. Suppose we denote this new domain as $\mathcal{F} \in 2^\Omega$. In order for the desired properties to hold we must also accept that $\mathcal{F}$ holds certain conditions.

Definition 1.4 (σ-algebra)**.** Let Ω be a set. We say that $\mathcal{F} \subset 2^\Omega$ is a σ-algebra (sometimes called a σ-field) of sets, 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 $\cup_{n=1}^\infty A_n \in \mathcal{F}$.

We can now formally define a probability space.

Definition 1.5 (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 $\mathbf{P}\colon \to [0, 1]$ is a probability function that satisfies:

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

In this case we shall call elements of $\mathcal{F}$ events.

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

Definition 1.6 (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 $A \cup B \in \mathcal{C}$.

We can immediately verify that any algebra $\mathcal{C}$ is closed under finite unions and finite intersections. We also notice that $\emptyset \in \mathcal{C}$, 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 1.1. If Ω is a set, and $A \subset \Omega$, then both 2^Ω and $\{\emptyset, A, A^c, \Omega\}$ are σ-algebras.

Example 1.2. 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 1.7 (Minimal Sigma Algebra)**.** Let Ω be a set, and let $H \subset 2^\Omega$ be a family of its subsets. Then we define the minimal sigma algebra that contains H, denoted $\sigma(H)$, as the intersection of all the σ-algebras that contains all the 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.

Example 1.3. (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:

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

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:

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

Note that in general, Borel's σ-algebra is defined to be the smallest σ-algebra that contains all the open sets in a general topological space. It can be showen that this definition is equivalent to the definitions we just gave for $\mathfrak{B}$ and $\mathfrak{B}_d$.

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

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

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

Now, if we consider again our previous problem, and let $\Omega = S^1$, in order to find a uniform probabiliy function on it we can 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. Now define $\widehat{\mathbf{P}}\colon \mathcal{G} \to [0, 1]$ as such:

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

We can see that $\widehat{\mathbf{P}}$ satisfies the conditions in Theorem 1.4 and thus exists an extension $\mathbf{P}$ defined on the sigma 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 we call $\mathbf{P}$ the uniform probability function on S^1.

Now we can more formally consider the properties of probability functions.

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

1. $\mathbf{P}(\emptyset) = 0$.
2. *If $\{A_n\}_{n=1}^N \subset \mathcal{F}$ are disjoint sets then $\cup_{n=1}^n A_n \in \mathcal{F}$ 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 $(B \setminus A) = \mathbf{P}(B) - \mathbf{P}(A)$ and thus $\mathbf{P}(A) \leq \mathbf{P}(B)$.*
5. *If $A, B \in \mathcal{F}$, then*
$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$$

Theorem 1.6 (Continuity of the probability function)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space.*

1. *If $(A_n)_{n=1}^\infty \subset \mathcal{F}$ is an increasing sequence of events, that is $A_1 \subset A_2 \subset A_3, \ldots$, then*
$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbf{A}_n.$$
2. *If $(A_n)_{n=1}^\infty \subset \mathcal{F}$ is a decreasing sequence of events, that is $A_1 \supset A_2 \supset A_3, \ldots$, then*
$$\mathbf{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbf{A}_n.$$

In fact the last proposition is a not more than a case of the following proposition.

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

Let us prove this theorem for the case $(A_n)_{n=1}^\infty$ is increasing. Define the following sequence:
$$B_1 = A_1$$
$$B_n = A_n \setminus A_{n-1}$$

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. $\cup_{n=1}^\infty B_n = \cup_{n=1}^\infty A_n$.

We now have:
$$\begin{aligned}\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbf{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty}(B_n) = \lim_{N\to\infty}\sum_{n=1}^{N}\mathbf{P}(B_n) = \lim_{N\to\infty}\mathbf{P}\left(\bigcup_{n=1}^{N} B_n\right)\\ = \lim_{N\to\infty}\mathbf{P}(A_N).\end{aligned}$$

2 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 probability of A given that B already happened as:

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

The intuition behind this definition should be clear. We calculate the probability of 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, let $B \in \mathcal{F}$ be an event such that $\mathbf{P}(B) > 0$. Then, the map $A \mapsto \mathbf{P}(A \mid B)$ is a probability function.*

The proof that the range of the function is $[0, 1]$ and that $(\Omega \mid B) = 0$ is clear from expanding the definitions, so we will only prove sigma additivity.

Proof. 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:

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

□

Proposition 2.2 (Law of Total Probability)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let $N \in \mathbb{N} \cup \{\infty\}$, and $(A_n)_{n=1}^{N}$ be disjoint events such that $\cup_{n=1}^{N} A_n = \Omega$. Then,*

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

Proof.

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

□

Example 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?

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. Thus,

$$\begin{aligned} \mathbf{P}(A_{n,k}) &= \mathbf{P}(A_{n,k} \cap (A_{n-1,k-1} \cup A_{n-1,k})) \\ &= \mathbf{P}(A_{n-1,k-1})\mathbf{P}(A_{n,k} \mid A_{n-1,k-1}) + \mathbf{P}(A_{n-1,k})\mathbf{P}(A_{n,k} \mid A_{n-1,k}). \end{aligned}$$

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.

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

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

and can be solved without much thought. Here's the general theoerm.

Theorem 2.3. (Bayes' Theorem). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let $N \in \mathbb{N} \cup \{\infty\}$, and $(A_n)_{n=1}^N$ be disjoint events such that $\cup_{n=1}^N A_n = \Omega$. Then,*

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

Proof. Left as an exercise to the reader. □

Example 2.2. Suppose we have a test for checking whether a person has the terrible the terrible "cooties". It has a true positive rate of 0.98, and a false positive rate of 0.01. Assume that 0.1% of the population has the cooties, what is the probability that a person who got a positive result has the cooties?

Denote,

$$\begin{aligned} A &= \{\text{the person is healthy}\} \\ B &= \{\text{the answer is positive}\}. \end{aligned}$$

From Bayes' theorem we have:

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

From the law of total probability we have

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

And thus,

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

3 Independance and Repeating Experiments

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

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

Thus we can use the definition of conditional probability to formally define Independence.

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

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

Notice that the interpretation that A is independent of B is only viable if we know that $\mathbf{P}(B) > 0$.

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. Clear from the definitions. □

Definition 3.2 (Independence)**.** 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 $\emptyset \neq K \subset \{1, 2, \ldots, N\}$ we have

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

Definition 3.3 (Pairwise Independence)**.** 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 $1 \leq i < j \leq N$ 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, 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.

Note that we only require independence for finite subsets and not for infinite subsets.

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 $(\widetilde{A}_n)_{n=1}^\infty$ such that $\widetilde{A}_n = A_n$ or $\widetilde{A}_n = A_n^c$ for each $n \in \mathbb{N}$. Then each choice of such $(\widetilde{A}_n)_{n=1}^\infty$ is (pairwise) independent.*

Proof. Using induction on the number of index such that we chose $\widetilde{A}_n = A_n^c$. To be completed. □

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

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space for a certain experiment. If we want to define a probability space for repeating the the experiment a finite number of times, which we will denote N, it makes sense to define it as such:

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

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

Similarly, when $N = \infty$, we will define the space to be:

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

Notice that we defined the probability function only on finite products of events. The extenstion to the rest of the sets will be done by Theorem 1.4.

Example 3.1 (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 defined as such:

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

Now for $N \in \mathbb{N} \cup \{\infty\}$ we have

$$\Omega_N = \{0, 1\}^N$$

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

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

$$\begin{aligned}
A_k &= \{k \text{ successes}\}. \\
H_i &= \{\omega_i = 1\}, \quad 1 \leq i \leq N.
\end{aligned}$$

We now notice that if $\omega = (\omega_1, \ldots, \omega_N) \in A_k$ then

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

where

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

From Theorem 3.2 and since the events are independent:

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

Finally, we get

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

Also, because we know that $(A_k)_{k=1}^N$ are all disjoint and that $\cup_{k=1}^N A_k = \Omega$ we get that:

$$\sum_{k=0}^{N} \P_N(A_k) = \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} = (p + (1-p))^N = 1,$$

just as expected.

Example 3.2 (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 just like the previous experiment, and indeed if we denote A_k the number of step the cat made to the right after in the product probability space $(\Omega_8, \mathcal{F}_8, \mathbf{P}_8)$, 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 step to the right would be $k-(8-k) = 2k-8$. Since we wonder about the probability of it being on the number 4, we should calculate the probability of A_6, and we will get that:

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

Notice that the result makes sense because as P increases we get higher results.

Example 3.3 (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

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

We may notice that the events $(A_n)_{n=1}^\infty$ are all independent, and also that

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

And as long as $p > 1$, we may assume that $\mathbf{P}_\infty(0,0,0,\dots) = 0$. Using that assumption, we have that

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

4 Random Variables

(add introduction)

Definition 4.1 (Random Variable)**.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A function $X\colon \Omega \to \mathbb{R}$ 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 shuold notice:

1. If we have that Ω is countable then as we know $\mathcal{F} = 2^{\Omega}$ and thus every function $X\colon \Omega \to \mathbb{R}$ 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 $X\colon \Omega \to \mathbb{R}$ a random variable. The distribution of X, denoted P_X is the function $\mathbf{P}_X\colon \mathfrak{B} \to [0,1]$ defined as such:

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

Remark 4.2. The space $(\Omega, \mathfrak{B}, \mathbf{P}_X)$ is a probability space.

Remark 4.3. Since a random variable X gives us data about experiments, for events $N \in \mathcal{F}$ such that $\mathbf{P}(N) = 0$ we don't care about the value of $X(N)$. From now on we won't define random variables for events with probability 0.

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

First we define our sample space

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

Since the balls are taken out uniformly we have

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

We now define the random variable:

$$\begin{aligned} &X\colon \Omega \to \mathbb{R} \\ &X((i,j,k)) = k \end{aligned}$$

and we notice that we want to calculate $\mathbf{P}(X \in \{17,18,19,20\})$. Since $\mathbf{P}_X$ is a probability function we can see that:

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

We notice that:

$$\mathbf{P}_X(\{k\}) = \frac{\binom{k-1}{2}}{\binom{20}{3}}$$

So finally we have:

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

Example 4.2 (Coupon Collector)**.** There are N types of coupons in a certain game, and we want to collect them all. Thus, each day we buy a new coupon, such that the probability of getting each one is the same. The sample space is $\Omega = \{1,2,\dots,N\}^{\mathbb{N}}$. We define a random variable $T\colon \Omega \to \mathbb{R}$ as such:

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

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 $j \geq N$ and $1 \leq j \leq N$ we define

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

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

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

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

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

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

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

And so get

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

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

$$\operatorname{supp}(f) := \operatorname{cl}_A\left(\{x \in A \,:\, f(x) \neq 0\}\right) = \overline{f^{-1}\left(\{0\}^{\mathrm{c}}\right)}.$$

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

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

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

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

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.

Proof. 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

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

□

The following theorem will show that F_X contains by itself all the information from P_X, but we will not prove it in this course.

Theorem 4.2. *Every CDF 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* $X\colon \Omega \to \mathbb{R}$, $Y\colon \Omega' \to \mathbb{R}$ *are two random variable, then*

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

Theorem 4.3. *Let* $F\colon \mathbb{R} \to [0,1]$ *be a function that satisfies all* 4 *basic properties of the CDF. Thus, exists a probability space* $(\Omega, \mathcal{F}, \mathbf{P})$ *and a random variable* $X\colon \Omega \to \mathbb{R}$ *such that*$F_X = F$.

We won't prove this theorem either, but using these two last theorems we can conclude that exists a bijection between all random variables' distributions and functions that satisfy the 4 properties from above.

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

Remark 4.4. Notice that if $X\colon \Omega \to \mathbb{R}$ is a random variable such that Ω is countable, then it is necessarily a discrete random variable.

Remark 4.5. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X\colon \Omega \to \mathbb{R}$ be a discrete random variable. Denote $\operatorname{supp}(X) = \{a_i\}_{i=1}^{\infty}$ and $p_i := \mathbf{P}(\{a_i\})$. By definition of the support we have that $\sum_{i=1}^{\infty} p_i = 1$, and thus, since $\mathbf{P}_X$ is a probability function we have that for any interval $(a,b) \subset \mathbb{R}$ we have

$$\mathbf{P}_X\left((a,b)\right) = \sum_{i\colon a_i \in (a,b)} p_i.$$

And that the CDF of X is

$$F_X(a)\mathbf{P}_X\left((-\infty, a)\right) = \sum_{i\colon a_i \le a} p_i.$$

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.

Definition 4.6 (Binomial Distribution)**.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X\colon \Omega \to \mathbb{R}$ 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 it's range is $\{0, 1, 2, \ldots, n\}$ and it's distribution satisfies

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

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

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

In this case we denote $X \sim \operatorname{Bin}(n,p)$.

When we talked about repeating Bernoulli experiments we saw that the number of "success" results in n experiments follows $\operatorname{Bin}(n,p)$ where p is the probability of success in a single experiment.

Example 4.3. 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 one or two?

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

Example 4.4. Suppose we flip a fair coin 2000 times. You may be able to convince yuor 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. The amount of times we get head follows $\text{Bin}(2000, 0.5)$ so we have

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

Using Stirling's approximation

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

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...

Definition 4.7 (Geometric Distribution)**.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X\colon \Omega \to \mathbb{R}$ be a random variable. We say that X is a geometric random variable with parameter $p \in [0, 1]$ if it's range is $\mathbb{N}$ and it's 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

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

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

Example 4.5. Suppose we had a Bernoulli experiment repeated indefininely, or until we get a success. Let p denote the probability of success. The random variable the calculates the first repetition in which the experiment succeeded follows $\text{Geo}(p)$.

Definition 4.8. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X\colon \Omega \to \mathbb{R}$ be a random variable. We say that X is a Poisson random variable with parameter $p > 0$ if it's range is $\mathbb{N}$ and it's 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 p if

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

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

This is a rather odd distribution. It is mostly used to model rare events. Suppose that an 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 a probability λ/n of succeeding. This bigger n we choose, the more accurate the model will get. As we take the limit $n \to \infty$, we obtain the Poisson distribution.

We can also notice that the Poisson distribution is really similar to the binomial distibution, in fact, it is exactly an approximation of it.

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

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

Proof.

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

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

$$\lim_{n\to\infty} ¶(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

□

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*

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

Before proving this theorem, we need cover some more things.

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 $m \in N$

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

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.

Finally, since every integer greater than 1 is divisible by at least one prime number

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

we get that

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

which completes the proof. □

Now that we have Euler's product formula we can prove Theorem 4.5.

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

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

Now from the exponent function properties we get that

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

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

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

Which proves that

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

and completes the proof. □

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

Remark 4.6. If X is a continuous random variable then for any $a \in \mathbb{R}$

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

Continuous random variables can get very weird, and in this course we will only discuss a very specific subset of random variables.

Definition 4.11 (Absolutely continuous random variable)**.** A random variable X is said to be absolutely continuous if exists an integrable function $f_X\colon \mathbb{R} \to [0, \infty)$ 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 of X, or PDF for short.

A couple of good points to notice are

1. From Theorem 4.2 we know that F characterizes P, which means that we can use f_X to calculate the probability of any event. For example let $A \subset \mathfrak{B}$ then

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

   but proving this equality is beyond the scope of these notes.

2. We always know that

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

Example 4.6. Let X be an absolutely continuous random variable with a probability density function f defined as such

$$f_X(y) = \begin{cases} C(2y - y^2), & y \in [0, 1] \\ 0, & \text{otherwise} \end{cases}.$$

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

To calculate C we can use the second fact we mentioned earlier

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

which implies that $C = \frac{3}{4}$. The probability that $X > 1$ is

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

Definition 4.12 (Uniform Distribution)**.** Let $[a, b] \subset \mathbb{R}$ be a closed interval. We say that a random variable X distributes uniformly on $[a, b]$ and denote $X \sim U([a, b])$ 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 distributes uniformly on $[a, b]$ if

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

Example 4.7. A bus reaches a train station every 15 minutes starting from 5 : 00. A student reachs the station in uniform distribution somewhere between 7 : 00 and 7 : 30. 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?

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

Definition 4.13 (Exponential Distribution)**.** We say that a random variable X distributes exponentially 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 distributes exponentially with parameter λ if

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

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 time events happen if the following conditions are satisfies

1. The events occur independently.
2. The rate of occurance is constant.
3. Two occurances can't happen at the same time.

We say that it is memoryless (elaborate later)

Definition 4.14 (Normal Distribution)**.** We say that a random variable X distributes normally with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, and denote $X \sim N(\mu, \sigma^2)$ 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 distributes normally 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.

In all the distributions we have seen so far, it was neccessary to show that $\int_{-\infty}^{\infty} f_X(y)\, dy = 1$. For most cases so far it's quite trivial to show that, but in the case of normal distribution we have some work to show. We will do that later. Also, 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 $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X \colon \Omega \to \mathbb{R}$ be a random variable. Then 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 parameter $0 \leq \alpha \leq 1$ such that*

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

Proof. To be added later. □

Definition 4.15 (Mixed Random Variable)**.** A random variable X that is not discrete or continuous, is called a mixed random variable. By the last theorem, we see that we can write it as a linear combination of a certain discrete and a certain random variables.

Remark 4.7. In a classic measure theory course it is standard to prove that every random variable can be written as a linear combination (with parameters that sum to 1) of a discrete variable, an absolutely continuous variable, and a singular continuous variable. Since we don't have the tools to discuss singular continuous variables yet we can ignore it in the bounds of this course.

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

Definition 4.16 (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 4.17 (Measurable function)**.** Let (X, Σ) and (Y, T) be measurable spaces. A fucntion $f\colon X \to Y$ is said to be measurable if for every $E \in \mathrm{T}$ the pre-image of E under f is in Σ; that is, for all $E \in \mathrm{T}$

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

That is, $\sigma(f) \subseteq \Sigma$, where $\sigma(f)$ is the σ-algebra generated by f. If f is a measurable function, one also may write

$$f\colon (X, \Sigma) \to (Y, \mathrm{T}).$$

Proposition 4.8. *Let $X\colon \Omega \to \mathbb{R}$ be a random variable, and $h\colon \mathbb{R} \to \mathbb{R}$ be Borel measurable, then $Y = h(X)$ is a random variable.*

No proof yet.

Theorem 4.9 (I don't remember who)**.** *Let $h\colon \mathbb{R} \to \mathbb{R}$ 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 points in $\mathbb{R}$ then h is measurable.*

Proof. In the analysis notes maybe? In measure theory for sure. □

Proposition 4.10 (Distribution of $Y = h(X)$)**.** *Let X be a random variable and h a measurable function.*

1. *If h is a strictly increasing function then $F_Y(a) = F_X(h^{-1}(a))$, where*

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

   if h is not a surjection. Additionaly, 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 $F_Y(a) = 1 - \lim_{y\to h^{-1}(a)^-} F_X(y)$, where*

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

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

Proof. Not today. □

The last propostion gave us a way to calculate the distribution of $Y = h(X)$ where h is strictly monotonic. Here is an example:

Example 4.8. Not now.

5 Expectation, Variance and Moments

The expectation of a random variable repesents the weighted "average" of its values.

Definition 5.1 (Expectation)**.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $X\colon \Omega \to \mathbb{R}$ be a random variable.

If X is a discrete random variable with $\operatorname{supp}(A) = \{a_i\}_{i=1}^{\infty}$ then we define

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

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 $[X]$ is undefined.

If X is an absolutely continuous variable with a probability density function f_X we define

$$\mathbf{E}_+[X] = \int_0^{\infty} y f_X(y)\, dy \qquad \mathbf{E}_-[X] = \int_{-\infty}^{0} y f_X(y)\, dy$$

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 $[X]$ is undefined.

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 $X \sim \operatorname{Bin}(n, p)$. Then

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

Expect more examples of expectation to come later.

We will see later that the average of results in n experiments as $n \to \infty$ is exactly $\mathbf{E}[X]$.

Proposition 5.1. *Let X be a nonnegative random variable. Then*

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

If X is a general random variable with a well defined expectation then

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

Proof. To be added. □

The following two propositions are known as the "Law of the unconscious statistician" or LOTUS for short.

Proposition 5.2 (LOTUS, discrete case)**.** *Let X be a discrete random variable with* $\operatorname{supp}(A) = \{a_i\}_{i\in\mathbb{N}}$*, and $h\colon \mathbb{R} \to \mathbb{R}$ be a measurable function. Then*

$$\mathbf{E}[h(X)] = \sum_{i\in\mathbb{N}} h(a_i)\mathbf{P}_X(b_i).$$

Proof. To be added. □

Proposition 5.3 (LOTUS, absolutely continuous case)**.** *Let X be an absolutely continuous random variable with a probability density function f_X. Suppose $h\colon \mathbb{R} \to \mathbb{R}$ is a piecewise monotonic function. Then*

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

This case is also true for continuous random variables in general.

Proof. To be added. □

Example 5.2. In a certain island, the population increases by 5% each year. In the first year, the popoulation 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?

We can notice that the number of years the population will survive follows follows a geometric distribution so we denote $N \sim \operatorname{Geo}(p)$. We know that the number of people on the island by the year is

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

Using the LOTUS formula for a discrete variable we get

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

And the final series is a geometric series so we get that

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

Example 5.3. Suppose a random square has a side length that is uniform on $[0,1]$. What is the expected area?

We denote $X \sim U([0,1])$, and then

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

Proposition 5.4 (Properties of Expectation)**.** *Let X be a random variable, $a \in \mathbb{R}$, and h_1, h_2 measurable functions.*

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* $\mathbf{P}(h_1(X) \geq a) = 1$, *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. To be added. □

Definition 5.2 (Variance and Standard Deviation)**.** Let X be a random variable with finite expectation. The Variance of X, denoted $\mathrm{Var}(X)$ is defined as

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

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

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

Intuitively, we say that $\mathrm{Var}(X)$ and σ_X describe how distant X is from its expectation. Another parameter that describes this is $\mathbf{E}\left[|X - \mathbf{E}[X]|\right]$ but we don't use it for a couple of reasons. Firstly, in many cases it's easier to caluclate the variance and standard deviation, and secondly, the standard deviation σ_X is the probabilistic version of Euclidean distance, and the variance $\mathrm{Var}(X)$ describes that square of the Euclidean distance, which are both very natural parameters.

Example 5.4. If we denote the result of a dice throw X, then we already know that $\mathbf{E}[X] = 3.5$ so the variance is

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

Insert here a table of variance by type of random variable.

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

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

From 1 *and* 2 *we get that* $\mathbf{E}[X]^2 \leq \mathbf{E}[X^2]$.

Proof. To be added. □

Definition 5.3 (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 as

$$\mu_n(X) = [X^n]$$

under the assumption it is well defined.

2. If X is a nonnegative random variable, then the $\alpha \in \mathbb{R}$ moment of X is defined as

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

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

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

In particular $C_1(X) = 0$ and $C_2(X) = \mathrm{Var}(X)$.

Example 5.5. Let $X \sim U([a,b])$, then

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

Add intuition and motivation for moments. Taylor expansion etc.

Definition 5.4 (Moment generating function)**.** Let X be a random variable. The moment generating function of X, denoted $M_X \colon \mathbb{R} \to [0, \infty]$ is defined by

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

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

Example 5.6. Let $X \sim \mathrm{Bin}(n,p)$, then

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

Proposition 5.6. *Let X be a random variable, and assume M_X is finite in $[-\epsilon, \epsilon]$ for some $\epsilon > 0$. Then M_X is a smooth function and for each $k \geq 0$*

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

Proof. To be added. □

Remark 5.1. Sometimes it's easier to calculate the moment generating function than the moments themselves. Using the previous proposition we can calculate the moments by differentiating M_X.

Moments have many uses, for example to find bounds for events of the form $\{X > a\}$ or $\{|X - \mathbf{E}[X]| > a\}$. The following proposition suggest another use.

Proposition 5.7. *Let X, Y be random variables. If 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 $\mathbf{P}_X = \mathbf{P}_Y$.*

6 Inequalities

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 $\operatorname{supp}(A) = \{a_i\}_{i\in\mathbb{N}}$ 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

$$|\mathbf{E}[X]| = \left|\int_{\mathbb{R}} y f_X(y)\, du\right| \leq \int_{\mathbb{R}} |y f_X(y)|\, dy = \int_{\mathbb{R}} |y|\, f_X(y) = \mathbf{E}[|X|].$$

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

Previously, we also saw that

$$\mathbf{E}[X]^2 \leq \mathbf{E}[X^2].$$

The last two inequalities are of the form

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

where h is the function $h(y) = |y|$ in the first inequality, and $h(y) = y^2$ in the second. We want to find for which functions h the inequality is satisfied

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

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

Some remarks may be good here.

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

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

Some more discussion about convex functions is required, then a short proof needs to be added.

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 $|X|$ has a moment β, that is $\mathbf{E}[|X|^\beta] < \infty$, then $\mathbf{E}[|X|^\alpha]$.*

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

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

□

Theorem 6.4 (Markov's inequality)**.** *Let X be a nonnegative random variable. 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

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

Because $g(y) \leq h(y)$ 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. □

Corollary 6.5. *If X is a random variable then for all $a > 0$*

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

More stuff

Theorem 6.6 (Chebyshev's inequality)**.** *Let X be a random variable with finite expectation. Then for every $a > 0$*

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

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

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

□

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

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

Theorem 6.7 (Weak law of large numbers, binomial case)**.** *Let $p \in [0, 1]$. For all $n \in \mathbb{N}$ denote $X_n \sim \mathrm{Bin}(n, p)$. Thus for any $\delta > 0$*

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

Proof. There are multiple. To be added later. □

The intuition behind this theorem is that it shows that for any margin $\delta > 0$ as we take $n \to \infty$ the number of successes divided by n is converges to p.

Theorem 6.8 (Weierstrass approximation theorem)**.** *Let $[a, b] \subset \mathbb{R}$ be a compact interval. Then, for every continuous function $f\colon [a, b] \to \mathbb{R}$ exists a polynomial sequence P_n that converges uniformly to f. In other words*

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

Proof. Better prove Stone-Weierstrass's theorem. □

7 Bertrand's Paradox

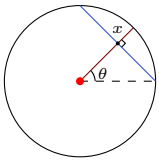
Before continuing to multivariate random variables, let's discuss Bertrand's paradox once again, and solve it. Recall the paradox:

> "Consider an equilateral triangle inscribed in a circle. Suppose a chord of the circle is chosen 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.

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 in perpendicular 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 on their respective intervals, and independently, 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

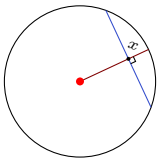
$$X\left((\theta, x)\right) = \text{the length of the chord corresponding to the pair } (\theta, x) = 2\sqrt{1 - x^2}.$$

Now we see that

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

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

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

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

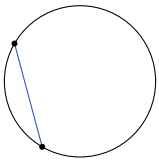
$$Y((x,y)) = \text{the length of the chord corresponding to the pair } (x,y) = 2\sqrt{1-x^2-y^2}.$$

Now we see that

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

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 should expect, calculating Z directly this time requires tedius geometric and trigonometric calculations. In the end of these calculations, we use similar methods to the ones shown in the first two solutions, and get

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

8 Multivariate Random Variables

Definition 8.1 (Multivariate random variable)**.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. The function $\mathbf{X} = (X_1, \ldots, X_N) \colon \Omega \to \mathbb{R}^N$ 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.

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

Proposition 8.1. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \ldots, X_N) \colon \Omega \to \mathbb{R}^N$. Then $\mathbf{X}$ is a multivariate random variable if and only if $X_k \colon \Omega \to \mathbb{R}$ 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

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

Now assume $X_1, \ldots, X_N$ 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 sigma algebra, and $X_1, \ldots, X_N$ are all random variables. □

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

$$\mathbf{P}_{\mathbf{X}}(A) = \mathbf{P}(\mathbf{X} \in A), \quad \forall A \in \mathfrak{B}_N.$$

Remark 8.1. 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 $\mathbf{P}_{X_1}, \ldots, \mathbf{P}_{X_N}$ because

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

The opposite, is not true.

Example 8.1. 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 X, and the number of white balls drawn Y. What is the probability of (X, Y)?

We see that

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

where i denotes the number of red balls drawn, and j the number of white balls drawn. From 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 (Cumulatice distribution function)**.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \ldots, X_N) \colon \Omega \to \mathbb{R}^N$ be a random vector. The cumulative ditribution function of $\mathbf{X}$, denoted $F_{\mathbf{X}}$, is the function $F_{\mathbf{X}} \colon \mathbb{R}^N \to [0,1]$ defined as

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

Proposition 8.2 (Properties of the CDF)**.**

1. *The function $F_{\mathbf{X}}$ is increasing in each coordinate.*
2. *The function $F_{\mathbf{X}}$ is continuous from the right in each coordinate.*
3. *For every choice $a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_N \in \mathbb{R}$ we have $\lim_{a_k \to -\infty} F_X(a_1, a_2, \ldots, a_N) = 0$*
4. *Denote $X^k = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_N)$ the random vector with the k-th coordinate erased. Then for every choice $a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_N \in \mathbb{R}$ we have*

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

In particular

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

Proof. To be added. □

Theorem 8.3. *Every CDF 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 $\mathbf{X} \colon \Omega \to \mathbb{R}^N$, $\mathbf{Y} \colon \Omega' \to \mathbb{R}^N$ are two random variable, then*

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

Theorem 8.4. *Let $F \colon \mathbb{R}^N \to [0,1]$ be a function that satisfies all 4 basic properties of the CDF. Thus, exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a random variable $\mathbf{X} \colon \Omega \to \mathbb{R}^N$ such that $F_{\mathbf{X}} = F$.*

We won't prove this theorem either, but using these two last theorems we can conclude that exists a bijection between all random variables' distributions and functions that satisfy the 4 properties from above.

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

Proposition 8.5. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $\mathbf{X} = (X_1, \ldots, X_N) \colon \Omega \to \mathbb{R}^N$ 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 $\mathbf{X} = (X_1, \ldots, X_N)$ is a discrete random vector. Then 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

$$1 \geq \mathbf{P}(X_1 \in B_k) = \mathbf{P}\left((X_1, X_2) \in B_k \times \mathbb{R}\right) \geq \mathbf{P}(\mathbf{X} \in B) = 1.$$

Thus we got that for all $1 \leq k \leq N$ that X_k is discrete. Now suppose that $X_1, \ldots, X_N$ are discrete random variables. Then exist countable sets $B_1, \ldots, B_N$ such that $(X_k \in B_k) = 1$ for all $1 \leq k \leq N$. Under the axiom of choice we have that the set $B = B_1 \times \cdots \times B_N$ is countable and it is clear that $\mathbf{P}(\mathbf{X} \in B) = 1$ as wanted which completes the proof. □

Remark 8.2. 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 $\mathbf{X} = (X_1, \ldots, X_N) \colon \Omega \to \mathbb{R}^N$ be a random vector. Then $\mathbf{X}$ is called an absolutely continuous random vector if $F_{\mathbf{X}}$ is continuous.

Remark 8.3. As in the case of a random variable, if $\mathbf{X} \colon \Omega \to \mathbb{R}^N$ is an absolutely continuous random vector, then $\mathbf{P}(\mathbf{X} = a) = 0$ for all $a \in \mathbb{R}^N$.

Similarly to general continuous random variables, we won't discuss general continuous random vectors in this course, and only consider absolutely continuous random vectors.

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

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

And for all $a \in \mathbb{R}^N$

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

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

Remark 8.4. Like in the case of random variables we konw that for any $A \subset \mathbb{R}^N$ such that $f \cdot \mathbb{1}$ is an integrable function

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

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

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

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

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

Example 8.2. Let (X, Y) be a random vector with a probability density function defined as

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

Find the value of c, and calculate $\mathbf{P}(X + Y \le 1)$.

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 thoerem

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

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

To find $\mathbf{P}(X+Y \leq 1)$, we need to understand that the function is only different than zero on $R = [0,1] \times [0,2]$. Which means that we need to find the integral of f_X on $T = R \cup \{(x,y) \mid x+y \leq 1\}$. 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 $\mathbf{X} = (X_1, \ldots, X_N) \colon \Omega \to \mathbb{R}^N$ be an absolutely continuous random vector with a probability density function $f_{\mathbf{X}} \colon \mathbb{R}^N \to [0,\infty)$. Then X_k is an absolutely continuous random variable for each $1 \leq k \leq N$, with a probability density function*

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

And the same is true for other sub random vectors.

Proof. To be added. □

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

$$\mathrm{Vol}(B) = \int_{\mathbb{R}^N} \mathbb{1}_B(x)\,dx$$

Definition 8.8 (Uniform distribution)**.** Let $\mathbf{X} \colon \Omega \to \mathbb{R}^N$ 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 denote $\mathbf{X} \sim U(B)$ if $\mathbf{X}$ is absolutely continuous with a density function

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

A theorem needs to be here.

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

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

Notes and Examples

Definition 8.10 (Measure)**.** Let (X, Σ) be a measurable space. A set function $\mu \to [0, \infty]$ 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

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

The second condition is called σ-additivity.

Definition 8.11 (Atom)**.** Given a measurable space (X, Σ), and a measure μ on that space, a set $A \subset X$ in Σ is called an atom if $\mu(A) > 0$ and if for any measurable subset $B \subset A$ then $0 \in \{\mu(B), \mu(A \setminus B)\}$.

Mixed random vectors.

Definition 8.12 (Convolution)**.** Let $f, g \colon \mathbb{R} \to [0, \infty)$. (In fact we can also use this definition for $f, g \colon \mathbb{R} \to \mathbb{R}$ such that $|f|, |g|$ are integrable). The convolution of f and g, denoted $f * g \colon \mathbb{R} \to [0, \infty)$ is the function

$$(f * g)(a) = \int_{\mathbb{R}} f(x) g(a - x)\, dx.$$

This is still missing a proof but

$$F_{X+Y} = f_X * F_Y = f_Y * F_X.$$

And

$$f_{X+Y} = f_X * f_Y.$$

9 Expectation, Covariance and Correlation

A natural extension of expectation to random vectors is to set

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

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

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

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

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

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

Definition 9.1. Let $\mathbf{X}\colon \Omega \to \mathbb{R}^N$ 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

$$\mathbf{E}[X] = (\mathbf{E}[X_1], \ldots, \mathbf{E}[X_n]).$$

Proposition 9.1. *Let $\mathbf{X}\colon \Omega \to \mathbb{R}^N$ be a random vector, and let $h\colon \mathbb{R}^N \to \mathbb{R}$ 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 $\operatorname{supp}(\mathbf{X}) = (\mathbf{a}^i)_{i \geq 1}$, then

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

Proof. To be added, possibly. □

Corollary 9.2 (Linearity of Expectation)**.** *Let $X, Y\colon \Omega \to \mathbb{R}$ be two random variables and $a, b \in \mathbb{R}$. Then*

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

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

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

This implies

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

But the converse is not always true.

For now, we won't write the proofs but here's a counter example to the previous direction of the above corollary.

Example 9.1. Let $X \sim U([-1,1])$ and $Y = X^2$. It is clear that X and Y are not independent, but we have

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

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

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

and thus immediately

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

Definition 9.2 (Uncorrelatedness)**.** Two random vectors $X, Y\colon \Omega \to \mathbb{R}$ are said to be uncorrelated if $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

It is clear that if X, Y are two independent random variables, then they are uncorrelated, but as we saw, the converse is not always true.

A general formula that follows directely from the definition of variance and the linearity of expectation is

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

Another such formula is

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

Intuitively, the last value represents how dependent X, and Y are. The farther it is from zero, the variables are less independent. Because it has multiple uses, we give it a special name.

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

$$\operatorname{Cov}(X,Y) = \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right] = \mathbf{E}[XY] - \mathbf{E}[X] - \mathbf{E}[Y].$$

From the definition of covariance we get the following convenient formula:

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

Additionaly, the variables X, Y are uncorrelated if and only if $\operatorname{Cov}(X,Y) = 0$.

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

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

Example 9.2. Suppose we made $n \geq 2$ experiments such that the probability to get 1 in the experiment is p_1, the probability to get 2 it p_2, and the probability to get 3 is $p_3 = 1 - p_1 - p_2$. We denote X the number of times we got 1, and Y the number of times we got 2 in the experiments. Intuitively, since as X increases there are less experiments to get 2 in Y should decrease. This should result in the covariance and correlation of X and Y to be negative. Indeed, since $X \sim \operatorname{Bin}(n, p_1)$ and $Y \sim \operatorname{Bin}(n, p_2)$ we get that

$$[X] = np_1 \quad \text{and} \quad [Y] = np_2.$$

We notice that

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

so 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 if $i = j$ the result can't be 1 and 2 so the expression inside the expectancy is zero, and thus the expectancy is also zero. If $i \neq j$ then the events are independent and we get

$$\begin{aligned}\mathbf{E}\left[\mathbb{1}_{\{\text{the } i \text{ result is 1}\}}\mathbb{1}_{\{\text{the } j \text{ result is 2}\}}\right] &= \mathbf{E}\left[\mathbb{1}_{\{\text{the } i \text{ result is 1}\}}\right]\mathbf{E}\left[\mathbb{1}_{\{\text{the } j \text{ result is 2}\}}\right] \\ &= \mathbf{P}(\text{the } i \text{ result is 1})\mathbf{P}(\text{the } j \text{ result is 2}) = p_1p_2\end{aligned}$$

Since there are $n(n-1)$ pairs (i, j) such that $i \neq j$ we get

$$\mathbf{E}[XY] = n(n-1)p_1p_2$$

Thus

$$\operatorname{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = n(n-1)p_1p_2 - np_1 \cdot np_2 = -np_1p_2.$$

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

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

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

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

Proof. The proofs for all of these properties follow from the definitions of variance, covariance, or the linearity of expectation. □

Notice that all random variables on a certain probability space $(\Omega, \mathcal{F}, \mathbf{P})$ forms a vector space. We can also notice that the only things that's preventing the map $(X, Y) \mapsto \operatorname{Cov}(X, Y)$ from defining an inner product, is that exist random variables $X \neq 0$ such that $\operatorname{Cov}(X, X) = 0$. Notably, all the constant random variables that are different than 0. There is a way to fix this, by requiring the expectations on all random variables to be 0. Now we can define the space

$$L_2(\Omega, \mathcal{F}, \mathbf{P}) = \{X \colon \Omega \to \mathbb{R} \mid X \text{ is a random variable}, \mathbf{E}[X] = 0, \operatorname{Var}(X) < \infty\}$$

coupled with the bilinear form $\langle X, Y\rangle := \operatorname{Cov}(X, Y)$. Then, the space $(L_2(\Omega, \mathcal{F}, \mathbf{P}), \langle\cdot,\cdot\rangle)$ is an inner product space, and moreover, it is a Hilbert space. By the Cauchy–Schwarz inequality we get that

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

for all $X, Y \in L_2(\Omega, \mathcal{F}, \mathbf{P})$. In fact, this inequality holds for any two random variables. If $\operatorname{Var}(X) = \infty$ or $\operatorname{Var}(Y) = \infty$ it holds very triviallly, so we may assume $\operatorname{Var}(X), \operatorname{Var}(Y) < \infty$. In this case, we know that $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ are both vectors in $L_2(\Omega, \mathcal{F}, \mathbf{P})$ and thus

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

Remark 9.1. The lasts proposition shows that the correlation of two random variables gives values in the interval $[0, 1]$. That is, for any $X, Y \colon \Omega \to \mathbb{R}$ we have $-1 \leq \rho_{X,Y} \leq 1$.

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

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

Proof. To be added. □

Example 9.3. Let $X \sim \operatorname{Bin}(n, p)$. Since all the experiments are independent we get that $X = \sum_{i=1}^{n} X_i$ for $X_i \sim \operatorname{Bin}(1, p)$ for $1 \leq i \leq n$. 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 are independent we can also easily calculate the variance:

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

Example 9.4. Let $\sigma \colon [n] \to [n]$ be a random permutation from S_n. Denote X the number of permutations of σ. Find $\mathbf{E}[X]$.

First we notice that

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

and thus

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

This method of using indicators to find the expectation is very useful and is sometimes called the "method of indicator".

10 Conditional Probability Function and Expectation

Definition 10.1 (Conditional probability function)**.** Let $X, Y \colon \Omega \to \mathbb{R}$ be two random variables and $a \in \mathbb{R}$ such that $\mathbf{P}_X(a) > 0$. The conditional probability of Y given that $X = a$ is denoted $\mathbf{P}_{Y|X}(\cdot \mid a) \colon \mathfrak{B} \to \mathbb{R}$ and defined as

$$\mathbf{P}_{Y|X}(B \mid a) = \mathbf{P}(Y \in B \mid X = a) = \frac{\mathbf{P}(Y \in B, X = a)}{\mathbf{P}(X = a)}$$

Here are some properties of the conditional probability function

1. $\mathbf{P}_{Y|X}(\cdot \mid a)$ is a probability function.
2. $\mathbf{P}_Y(b) = \sum_a \mathbf{P}_X(a)\mathbf{P}_{Y|X}(b \mid a)$ for all $b \in \mathbb{R}$.
3. If $\mathbf{P}_Y(b) > 0$, then
$$\mathbf{P}_{Y|X}(a \mid b) = \frac{\mathbf{P}_{Y|X}(b \mid a)\mathbf{P}_X(a)}{\mathbf{P}_Y(b)}.$$

Example 10.1. Let X be the number of successes in N independent Bernoulli experiments with probability p of success, such that $N \sim \text{Poi}(\lambda)$. Find the distribution of X.

We get that for $k \geq 0$

$$\begin{aligned}
\mathbf{P}_X(k) &= \sum_{n=0}^{\infty} \mathbf{P}_N(n)\mathbf{P}_{X|N}(k \mid n) = \sum_{n=k}^{\infty} \mathbf{P}_N(n)\mathbf{P}_{X|N}(k \mid n) \\
&= \sum_{n=k}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!} \cdot \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{e^{-\lambda}}{k!} p^k \lambda^k \sum_{n=k}^{\infty} \frac{\lambda^{n-k}}{(n-k)!}(1-p)^{n-k} \\
&= \frac{e^{-\lambda}}{k!} p^k \lambda^k \sum_{n=0}^{\infty} \frac{\lambda^n}{(n)!}(1-p)^n = \frac{e^{-\lambda}}{k!} p^k \lambda^k e^{\lambda(1-p)} = e^{-\lambda p}\frac{(\lambda p)^k}{k!}.
\end{aligned}$$

Add discussion about more than one variable

Definition 10.2 (Conditional expectation for an event)**.** Let $X, Y \colon \Omega \to \mathbb{R}$ be two discrete random variables and $a \in \mathbb{R}$ such that $\mathbf{P}_X(a) > 0$. The condition expectation of Y given that $X = a$ is

$$\mathbf{E}[Y \mid X = a] = \sum_b b\mathbf{P}_{Y|X}(b \mid a).$$

Example 10.2. To be added.

Definition 10.3 (Conditional expectation, discrete variables)**.** Let $X, Y \colon \Omega \to \mathbb{R}$ be two discrete random variables. Suppose that $\text{supp}(X) = \{a_i\}_{i \geq 1}$. The conditional expectation of Y given X is a random variable $\mathbf{E}[Y \mid X] \colon \Omega \to \mathbb{R}$ defined as

$$\mathbf{E}[Y \mid X](\omega) = \begin{cases} \mathbf{E}[Y \mid X = a_1], & \text{if } X(\omega) = a_1 \\ \mathbf{E}[Y \mid X = a_2], & \text{if } X(\omega) = a_2 \\ \mathbf{E}[Y \mid X = a_3], & \text{if } X(\omega) = a_3 \\ \vdots \end{cases}$$

Here are some properties of conditional expectation

1. To be added

More stuff

Definition 10.4 (Conditional PDF)**.** a

Definition 10.5 (Conditional expectation, absolutely continuous variables)**.** Let.

11 Borel–Cantelli Lemmas

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and $(A_n)_{n\geq 1}$ an event sequence in $\mathcal{F}$. We denote

$$\limsup_{n\to\infty} A_n = \{\omega \in A_n \text{ infinitely often}\} = \{\omega \in A_n \text{ i.o.}\}.$$

Similarly,

$$\liminf_{n\to\infty} A_n = \{\omega \in A_n \text{ eventually}\}.$$

It is clear that

$$\liminf_{n\to\infty} A_n \subset \limsup_{n\to\infty} A_n.$$

Definition 11.1. Let $(A_n)_{n\geq 1}$ be an event sequence. Then A_n is said to have a limit if

$$\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n.$$

In this case, we denote

$$\lim_{n\to\infty} A_n = \liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n.$$

Here add examples.
Using this definition can now generalize Theorem 1.6.

Theorem 11.1 (Continuity of the probability function)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(A_n)_{n\geq 1}$ an event sequence in $\mathcal{F}$ such that $\lim_{n\to\infty} A_n$ exists. Then,*

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

To prove this theorem, we need the following lemma

Lemma 11.2 (Fatou's lemma)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(A_n)_{n\geq 1}$ an event sequence in $\mathcal{F}$. Then*

1. $\mathbf{P}(\liminf_{n\to\infty} A_n) \leq \liminf_{n\to\infty} \mathbf{P}(A_n)$
2. $\mathbf{P}(\limsup_{n\to\infty} A_n) \geq \limsup_{n\to\infty} \mathbf{P}(A_n)$

Proof. later □

We will know prove Theorem 11.1.

Proof. later □