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 $\Omega = \{1,2,\dots,11,12\}$ for the sum of the results of the dice.

Definition 1.2 (Probability space, intuitive definition)**.** A discrete probability space is a pair $(\Omega, \mathbf{P})$, where Ω is a countable sample set, and $\mathbf{P}\colon \Omega \to [0,1]$ 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 called 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*
$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$$

Proposition 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:
$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbf{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} (B_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mathbf{P}(B_n) = \lim_{N\to\infty} \mathbf{P}\left(\bigcup_{n=1}^{N} B_n\right)$$
$$= \lim_{N\to\infty} \mathbf{P}(A_N).$$

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 $\emptyset \neq K \subset \{1, 2, \dots, N\}$ we have

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

Definition 3.3 (Pairwise 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:

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

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 $\omega = (\omega_1, \dots, \omega_N) \in A_k$ 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:

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

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:

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

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 called a random variable if for any open interval $(a,b) \subset \mathbb{R}$ we have

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

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

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

We now define the random variable:

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

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

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

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 $\Omega = \{1, 2, \ldots, N\}^{\mathbb{N}}$. We define a random variable $T\colon \Omega \to \mathbb{R}$ as such:

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

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

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

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

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

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

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

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. 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 $\{0,1,2,\dots,n\}$ and it's distribution satisfies

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

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

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

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 distributes $\operatorname{Bin}(n,p)$ where p is the probability of success in a single experiment.

Example 4.3. For example, 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 distributes $\text{Bin}(2000, 0.5)$ 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

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

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