Math Notes

Tyler Wilson

Contents

1 Probability

1.1 Foundations

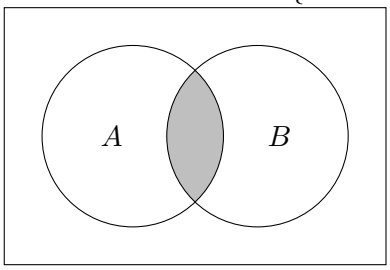
1.1.1 Notation

We can define the following notation for probability:

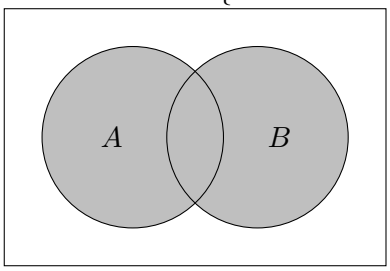
- We denoted the *state space* or *universal set* as Ω or S.
- An *event* is a "nice" subset of the state space, i.e. an event A fulfils $A \subseteq \Omega$. The set of all events is often denoted by $\mathcal{A}$, $\mathcal{E}$, or $\mathcal{F}$.
- A *probability measure* $\mathbb{P}$ is a map taking events as argument with
  - $\mathbb{P}: \mathcal{A} \to [0,1]$
  - $\mathbb{P}(\Omega) = 1$
  - For a countable index set I with $(A_i)_{i\in I}$ being disjoint events, we have $\mathbb{P}\left(\bigsqcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i)$

Set notations:

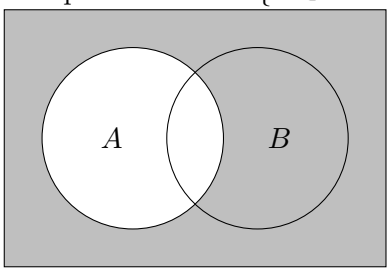
- Intersection: $A \cap B = \{\omega \in \Omega : \ \omega \in A \text{ and } \omega \in B\}$



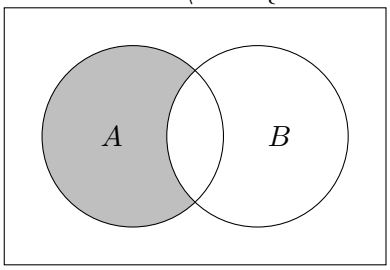
- Union: $A \cup B = \{\omega \in \Omega : \ \omega \in A \text{ or } \omega \in B\}$



- Complement: $A^c = \{\omega \in \Omega : \ \omega \notin A\}$



- Difference: $A \setminus B = \{\omega \in A : \ \omega \notin B\}$



We use the square cups $\sqcup$ to denote the disjoint union of sets. This means that $A_1 \sqcup A_2$ implies that the two sets have no overlap and they are disjoint: $A_1 \cap A_2 = \emptyset$.
Discrete Uniform Distribution:
We require that Ω is non empty and finite. If every event ω is equally likely, we have a discrete uniform distribution. This means that

$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} \ \forall \omega \in \Omega$$

Then the probability of an event A is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

We call this setting a discrete uniform distribution or uniformly random.
Ex: For a die, we have $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\mathbb{P}(\{\omega\}) = \frac{1}{6}$.
Ex2: For two die we have $\Omega = \{(1, 1), (1, 2), \ldots, (6, 5), (6, 6)\}$ and the probability of rolling a 1, 2, or 3 is

$$\mathbb{P}(\{1, 2, 3\}) = \mathbb{P}(\{1\} \sqcup \{2\} \sqcup \{3\}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{2\}) + \mathbb{P}(\{3\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Some useful properties of probability:
For disjoint events $A,\ B$ we have $\mathbb{P}(A \sqcup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
Let $A,\ B$ be events with $A \subset B$. Then $\mathbb{P}(A) \leq \mathbb{P}(B)$

Proof. Define $C := B \setminus A$. By definition we have $\mathbb{P}(C) \geq 0$, $B = A \sqcup C$ and hence

$$\mathbb{P}(B) = \mathbb{P}(A \sqcup C) = \mathbb{P}(A) + \mathbb{P}(C) \geq \mathbb{P}(A) + 0 = \mathbb{P}(A)$$

□

For any event A it holds that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

Proof. Observe that $\Omega = A \sqcup A^c$. Hence

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$$

□

Let A and B be two events. Then

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

Proof. Observe that $A \cup B = (A \setminus B) \sqcup (A \cap B) \sqcup (B \setminus A)$ which is a disjoint union. Hence,

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) \\
&= \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) \\
&= \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}((A \setminus B) \sqcup (A \cap B)) + \mathbb{P}((A \cap B) \sqcup (B \setminus A)) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)
\end{aligned}$$

□

This is known as the inclusion-exclusion principle. The general case for $n \geq 2$ events is

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right)$$

Proof. We begin by assuming the above hypothesis is true for $n \geq 2$.
We will also assume that the equation $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ is taken as an axiom.
The base case of $n = 2$ gives

$$\mathbb{P}\left(A_1 \cup A_2\right) = \sum_{k=1}^{2} (-1)^{k+1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right)$$
$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

which is the same as our axiom, thus the base case is true.
For $n + 1$ case we can write

$$\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \mathbb{P}\left(\left(\bigcup_{i=1}^{n} A_i\right) \cup A_{n+1}\right)$$

We can then use the base case to rewrite this as

$$\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) + \mathbb{P}(A_{n+1}) - \mathbb{P}\left(\left(\bigcup_{i=1}^{n} A_i\right) \cap A_{n+1}\right)$$

The last term in our expression can be rewritten using the distributive law as

$$\mathbb{P}\left(\left(\bigcup_{i=1}^{n} A_i\right) \cap A_{n+1}\right) = \mathbb{P}((A_1 \cap A_{n+1}) \cup \cdots \cup (A_n \cap A_{n+1})) = \mathbb{P}\left(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})\right)$$

And so we have

$$\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) + \mathbb{P}(A_{n+1}) - \mathbb{P}\left(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})\right)$$

The first and last terms are now unions of n for which we assumed the formula to hold.
Expanding these out gives

$$\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \left(\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) - \mathbb{P}\left(\left(\bigcap_{j=1}^{k} A_{i_j}\right) \cap A_{n+1}\right)\right) + \mathbb{P}(A_{n+1})$$

Combining these terms gives

$$\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \sum_{k=1}^{n+1} (-1)^{k+1} \sum_{1 \leq i_1 < \ldots < i_k \leq n+1} \mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right)$$

Which justifies the inductive step for $n + 1$. □

1.1.2 Repeated Experiments

Consider the experiment where we have a jar of n balls, numbered 1 to n. If we take a ball out of the jar, there are n possible outcomes.
If we repeat the experiment and put the ball back in the jar, the next draw is independent of the previous one. This is known as *selection with replacement*. The number of possible outcomes will then be n^2. If this experiment is repeated k times then the number of possible outcomes will be n^k.
Now let's consider the case where we don't put the ball back in the jar. This is known as the case of *selection without replacement*. The number of ways to arrange n objects is

$$n(n-1)\cdots 2\cdot 1 = n!$$

Now, if we only select k balls, the number of possible outcomes is

$$n(n-1)\cdots(n-(k-1)) = \frac{n!}{(n-k)!}$$

So far we have been looking at the case where order matters. Some cases wherre the order of events matters to us when we are concerned about events happening sequentially, such as if you roll one dice and then another one. Basically any time when the position of an event is a consideration such as a time or belongs to a specific physical quantity. Cases where the order doesn't matter are if we have an unordered set for example. Say we want to select a committee of 3 people from a group of 10. The order in which we select the people doesn't matter.
In the case where order doesn't matter, we have the following formulae:
Selection with replacement:

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$

Selection without replacement:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

We get this formula by dividing the ordered case by the number of ways to order the k objects.
Ex: Given a 52 card deck, what is the probability of getting a flush of hearts?
We have 52 cards in the deck and 13 of them are hearts. We want to select 5 cards. We can define the event of getting a flush of hearts as A. We can then write

$$\Omega = \{(\omega_1,\ldots,\omega_5) :\ \omega_i \in \{1,\ldots,52\},\ \omega_i \neq \omega_j,\ i \neq j\}$$
$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{52}$$

This resembles sampling without replacement. We can then write

$$|\Omega| = \frac{52!}{(52-5)!}$$
$$A = \{(\omega_1,\ldots,\omega_5) \in \Omega :\ \omega_i \in \{1,\ldots,13\}\}$$

$$|A| = \frac{13!}{(13-5)!}$$
$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{13!}{(13-5)!} \cdot \frac{(52-5)!}{52!} = \frac{33}{66640} \approx 0.000495$$

Ex2: If we roll two dice, what is the probability of getting different numbers?
We can define the event of getting different numbers as A. We can then write

$$\Omega = \{(\omega_1, \omega_2) : \ \omega_1, \omega_2 \in \{1, \ldots, 6\}\}$$
$$|\Omega| = 6^2 = 36$$
$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{36}$$
$$A = \{(\omega_1, \omega_2) \in \Omega : \ \omega_1 \neq \omega_2\}$$
$$|A| = \frac{6!}{(6-2)!} = 6 \cdot 5 = 30$$
$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{30}{36} = \frac{5}{6} \approx 0.833$$

Ex3: In a classroom are 23 people. Assume that each one of them is independently equally likely to have their birthday on each of the 365 days of the year (yes, we also assume 365 for each year). Calculate the probability of at least two people having the same birthday.

Let us represent the birthday of one person as $\omega_i \in \{1, 2, \ldots, 365\}$. The universal set, Ω, can be formed as an ordered set of the birthdays of the 23 people:

$$\Omega = \{(\omega_1, \ldots, \omega_{23}) : \ \omega_1, \ldots, \omega_{23} \in \{1, \ldots, 365\}\}$$

Because multiple people can have the same birthday it resembles the problem of sampling with replacement. The cardinality will then be

$$|\Omega| = n^k = 365^{23}$$

We can define the event A to be that at least two people share a birthday. The compliment of the problem A is A^c and is that everyone has a unique birthday. Given the probability of the compliment is easier to compute than the probability of the event we can compute the probability of the event as

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$$

For everyone to have a unique birthday resembles the problem of sampling without replacement as we cannot reuse the same numbers (can't have the same birthday twice).

$$|A^c| = \frac{n!}{(n-k)!} = \frac{365!}{(365-23)!}$$

The probability of the event is then

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \frac{|A^c|}{|\Omega|} = \frac{\frac{365!}{(365-23)!}}{365^{23}} = \frac{365!}{365^{23}342!}$$

$$\mathbb{P}(A) = 1 - \frac{364!}{365^{22}342!} = 1 - \frac{1}{365^{22}} \prod_{i=343}^{364} i$$

$$\mathbb{P}(A) \approx 0.507$$

Ex4: I have three non-standard six-sided dice. The red one has five sides showing 4 and one side showing 1. The green one has three sides showing 2 and three showing 5. The blue one has five sides showing 3 and one side showing 6.
Consider a game where the first player chooses one of the dice and then the second player then selects one from the two that remain. Both players roll their dice and the player with the highest score wins. Do you want to be the first player or second player in this game?
(You could start by using the method in the "table" example to compute the winning probabilities for each of the three possible pairs of dice.)

Because of the nonordinary arrangement of the dice, the easiest way to solve this problem may be to write out each possibility. We can split this problem into three distinct cases:

1. Red vs. green
2. Red vs. blue
3. Green vs. blue

Looking at the first case, if we have one roll with the red die and one roll with the green die we can make a table of each possible combination.

possibilities	G2	G5
R4	$5 \cdot 3$	$5 \cdot 3$
R1	$1 \cdot 3$	$1 \cdot 3$

The cases where green wins are boxed in green and the cases where red wins are boxed in red. Summing the total possibilities we get that red wins in $\frac{15}{36}$ cases and green wins in $\frac{21}{36}$ cases so green is favored to win with a probability of $\frac{7}{12}$.
Let us repeat this with the other two cases now:
Red vs. blue:

possibilities	B3	B6
R4	$5 \cdot 5$	$5 \cdot 1$
R1	$1 \cdot 5$	$1 \cdot 1$

Here we get that red wins in $\frac{25}{36}$ cases and blue wins in $\frac{11}{36}$ cases so red is favored to win.
Green vs. blue:

possibilities	B3	B6
G2	$3 \cdot 5$	$3 \cdot 1$
G5	$3 \cdot 5$	$3 \cdot 1$

Here we get that blue wins in $\frac{21}{36}$ cases and green wins in $\frac{15}{36}$ cases so blue is favored to win with a probability of $\frac{7}{12}$.

The game resembles that of rock paper scissors in that one of the dies is favored to beat one of the other two but lose to the other. Given this, it makes more sense to choose second. This way you can see what the other person selects and then choose the die that is favored to beat it. In the event that you have to choose first, selecting green or red gives the best odds of winning.

1.1.3 Conditional Probability

Ex: Roll a D6. If there's a dot in the lower right corner, what is the probability of it being an odd number?
We can define the event of rolling an odd number as B and the event of having a dot in the lower right corner as A. We can then write

$$A = \text{dot in corner} = \{4, 5, 6\}$$
$$B = \text{odd number} = \{1, 3, 5\}$$
$$A \cap B = \{5\}$$
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$$
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

Note that $\mathbb{P}(B|A)$ is the probability of B given A. It is defined as

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

Ex2: If you roll two D6 and the sum is 8, what is the probability that a 6 was rolled?

$$\Omega = \{(\omega_1, \omega_2) : \ \omega_1, \omega_2 \in \{1, 2, 3, 4, 5, 6\}\}$$
$$A = \{(\omega_1, \omega_2) : \omega_1 + \omega_2 = 8\} = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$$
$$B = \{(\omega_1, \omega_2) : \omega_1 = 6 \text{ or } \omega_2 = 6\} = \{(1, 6), (2, 6), (3, 6), (4, 6), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5)\}$$
$$A \cap B = \{(2, 6), (6, 2)\}$$
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{2/36}{5/36} = \frac{2}{5}$$

Ex3: In a deck of cards what is the probability that you draw the Queen of Spades followed by the King of Hearts and then the 10 of Clubs?

$$\Omega = \{(\omega_1, \omega_2, \omega_3) : \ \omega_i \in \{1, \ldots, 52\}\}$$

A is we draw the Queen of Spades first
B is we draw the King of Hearts second
C is we draw the 10 of Clubs third

$$\mathbb{P}(A) = \frac{1}{52}$$
$$\mathbb{P}(B|A) = \frac{1}{51}$$
$$\mathbb{P}(C|A \cap B) = \frac{1}{50}$$
$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B|A)\mathbb{P}(C|A \cap B) = \frac{1}{52}\frac{1}{51}\frac{1}{50}$$

In general,

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\cdots\mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1})$$

For two sets,

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

The Law of Total Probability:
If we partition Ω into n disjoint events then we can express the probability of some event B as

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

One common partition of Ω is $\Omega = A \sqcup A^c$. This leads to

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)$$

Bayes' Theorem:
We can express Bayes' Theorem by starting with the expression for conditional probability.

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

Combining this with the Law of Total Probability we can get the more general expression for Baye's Theorem:

$$\mathbb{P}(A_k|B) = \mathbb{P}(B|A_k)\frac{\mathbb{P}(A_k)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}, \ \forall k \in \{1, \ldots, n\}$$

such that Ω is split up into n disjoint partitions, $A_1, \ldots, A_n$. Ex: Celiac occurs in about $\frac{1}{135}$ people (C). Say about 5% of people have similar symptoms (C) and don't have it. Also assume that the probability that you have symptoms if you have celiac is 99%. What is the probability of having celiac if you have the symptoms?

$$\mathbb{P}(C) = \frac{1}{135}$$
$$\mathbb{P}(S|C^c) = 0.05$$
$$\mathbb{P}(S|C) = 0.99$$
$$\mathbb{P}(C|S) = \mathbb{P}(S|C)\frac{\mathbb{P}(C)}{\mathbb{P}(S)}$$
$$\mathbb{P}(S) = \mathbb{P}(S|C)\mathbb{P}(C) + \mathbb{P}(S|C^c)\mathbb{P}(C^c)$$

$$\mathbb{P}(C|S) = \frac{\mathbb{P}(S|C)\mathbb{P}(C)}{\mathbb{P}(S|C)\mathbb{P}(C) + \mathbb{P}(S|C^c)\mathbb{P}(C^c)} = \frac{0.99 \cdot \frac{1}{135}}{0.99 \cdot \frac{1}{135} + 0.05 \cdot \frac{134}{135}} \approx 0.13$$

Two events A and B are independent if

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

Ex: Roll two D6. Let A be the event that the first die is a 6 and B be the event that the second die is a 6.

$$\mathbb{P}(A) = \frac{1}{6}$$
$$\mathbb{P}(B) = \frac{1}{6}$$
$$\mathbb{P}(A \cap B) = \frac{1}{36}$$
$$\mathbb{P}(A)\mathbb{P}(B) = \frac{1}{36} = P(A \cap B)$$

So A and B are independent.
Ex2: Roll two D6. Let A be the event that the sum is 5 and B be the event that there is at least one 1 rolled.

$$A = \{(1,4),(2,3),(3,2),(4,1)\}$$
$$B = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(3,1),(4,1),(5,1),(6,1)\}$$
$$A \cap B = \{(1,4),(2,3),(3,2),(4,1),(1,1)\}$$
$$|A| = 4,\ |B| = 11,\ |A \cap B| = 2$$
$$\mathbb{P}(A \cap B) = \frac{2}{36}$$
$$\mathbb{P}(A) = \frac{4}{36}$$
$$\mathbb{P}(B) = \frac{11}{36}$$
$$\mathbb{P}(A)\mathbb{P}(B) = \frac{44}{1296} \neq \frac{2}{36}$$

So A and B are not independent.

1.2 Random Variables

A random variable is a measurable function on the sample space. Usually, $X : \Omega \to \mathbb{R}$
We usually use capital letters from the end of the alphabet to denote random variables.
They are a great way to no longer have to care about the exact probability space.
Ex: $\Omega = \{H, T\}$, $\mathbb{P}(\{H\}) = \frac{1}{2} = \mathbb{P}(\{T\})$
So $X(H) = 1$, $X(T) = 0$
$(\Omega, \mathbb{P})$ is the probability space.
$\mathbb{P}(X = 1) = \mathbb{P}(X = 0) = \frac{1}{2}$

1.2.1 Discrete Probability Distributions

Discrete uniform random variable:
$\mathbb{P}(X = k) = \frac{1}{n}$ for all $k \in \{1, \dots, n\}$
This is denoted by $X \sim \text{Unif}(\{1, \dots, n\})$
For example a toss of a coin can be represented as $X \sim \text{Unif}\{0, 1\}$ and the roll of a D6 can be represented as $X \sim \text{Unif}\{1, 2, 3, 4, 5, 6\}$
Bernoulli random variable:
$\mathbb{P}(X = 1) = p$ where $p \in [0, 1]$ and $\mathbb{P}(X = 0) = 1 - p = q$
Some examples of this include a coin toss (with a fair or unfair coin) and the probability of the answer to a yes/no question. Basically any trial where you can get either a true or false answer.
Binomial random variable:
This represents k number of successes in n independent Bernoulli trials.

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Represented by $X \sim \text{Bin}(n, p)$ or $X \sim \text{Binom}(n, p)$.
For example, occurances of heads for a coin or number of passengers that miss a flight.
Ex: Throw a coin n times. What is the probability ofo getting at least 2 heads?

$$\begin{aligned}
\mathbb{P}(X \geq 2) &= 1 - \mathbb{P}(\{x \geq 2\}^c) = 1 - \mathbb{P}(\{X < 2\}) \\
&= 1 - \mathbb{P}(X = 1) - \mathbb{P}(X = 0) \\
&= 1 - \binom{n}{1} p(1-p)^{n-1} - \binom{n}{0} (1-p)^n, \quad p = \frac{1}{2} \\
&= 1 - n\left(\frac{1}{2}\right)^n - \left(\frac{1}{2}\right)^n \\
&= 1 - (n+1)\left(\frac{1}{2}\right)^n
\end{aligned}$$

Probablity mass functions (pmf):
Let X be a discrete random variable and $(x_i)_{i \in I}$. The sequence of all values X can adopt, $p(x_i) = \mathbb{P}(X = x_i) \to p$ is the probability mass function of X. Note that

$$\sum_{i \in I} p(x_i) = 1$$

To test this with the Bernoulli random variable:

$$p(0) + p(1) = (1 - p) + p = 1$$

Discrete uniform:

$$\sum_{k=1}^{n} p(k) = \sum_{k=1}^{n} \frac{1}{n} = \frac{n}{n} = 1$$

Binomial:

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} 1^k = (p + (1-p))^n = 1^n = 1$$

Geometric Distribution:
Ex: Roll a die until you get a 6 (do an experiment until success).
We can represent this type of expirement as a binomial distribution with $n = \infty$ and p representing the probability of a successful experiment and $1 - p$ the probability of an unsuccessful experiment.

$$p = \frac{1}{6}$$

$$\sum_{k=1}^{\infty} p(1-p)^k = p \sum_{k=0}^{\infty} (1-p)^{k-1+1}$$

$$\sum_{k=0}^{\infty} a^k = \frac{1}{1-a}$$

$$p \sum_{k=0}^{\infty} (1-p)^{k-1+1} = p \frac{1}{1-(1-p)} = 1$$

1.2.2 Continuous Probability Distributions

We say that a random variable, X has a continuous distribution if there exists a function $f : \mathbb{R} \to \mathbb{R}$ such that for any "nice" $b \subseteq \mathbb{R}$

$$\mathbb{P}(X \in B) = \int_B f_X(x) dx$$

We call f_X the probablility density function (pdf) of X.
We will also get from this definition that

$$f_X(x) \geq 0 \ \forall x \in \mathbb{R}$$

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

For any random variable, X, we define its cumulative distribution function (cdf) F_X for all $x \in \mathbb{R}$ by

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(-\infty < X \leq x) = \int_{-\infty}^{x} f_X(y) dy$$

We also have that

- F_X is non-decreasing
- $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$
- $F'_X(x) = f_X(x)$

For a regular continuous probability distribution, it is worth noting that the probability of a singular point is 0.

$$\mathbb{P}(a < X \leq b) = \int_a^b f_X(x)dx$$
$$\mathbb{P}(X \leq b, X > a) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a)$$

Now if we say $b = a$ then we get

$$\mathbb{P}(X = a) = \mathbb{P}(a \leq X \leq a) = \int_a^a f_X(x)dx = 0$$

So we ignore all singular points. This is distinctly different than the discrete case where all we care about is singular points.

Continuous Uniform Distribution:

This is the case where the pdf is constant on some interval $[a, b]$ and 0 elsewhere.

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{a} 0dx + \int_a^b cdx + \int_b^{\infty} 0dx = c(b-a) = 1$$
$$\Rightarrow c = \frac{1}{b-a}$$

We say that X is uniformly distributed on the interval $[a, b]$, for $a < b$ if its pdf is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}$$

We write $X \sim U([a, b])$ or $X \sim U[a, b]$ or $X \sim \text{Unif}([a, b])$ The cdf is given by

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

Ex: if $T \sim U([1, 5])$ what is $\mathbb{P}(T \geq 2.5)$

$$\mathbb{P}(T \geq 2.5) = \int_{2.5}^{\infty} f_X(x)dx = \int_{2.5}^{5} \frac{1}{4}dx = \frac{5 - 2.5}{4} = \frac{2.5}{4} = \frac{5}{8}$$

Note: Anoter way we can represent the uniform distribution that avoids the piecewise notation is using the indicator function. The indicator function is defined as

$$\mathbb{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

So we can write the pdf as $f_X(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$

Exponential Distribution:

Ex: Radioactive decay.

Radioactive atoms spontaneously decay. If the atom has not yet decayed, then the probability that it decays in the next short time Δt is $\approx \lambda \Delta t$ for a constant $\lambda > 0$. It does not matter how long the atom has been around. If it has not decayed up to now then the probability now is still $\approx \lambda \Delta t$. We can express this as

$$\mathbb{P}(X \in [0, \Delta t]) \approx \lambda \Delta t$$
$$\mathbb{P}(X \in [t, t+\Delta t] | X \geq t) = \mathbb{P}(X \in [0, \Delta t])$$

This is an example of a *memoryless process* as the probability of decay is independent of the time the atom has been around (it doesn't care about the events that happened before). The memoryless property can be represented as $\mathbb{P}(X \geq t+s | X \geq t) = \mathbb{P}(X \geq s)$.

$$\mathbb{P}(X \in [t, t+\Delta t] | X \geq t) = \frac{\mathbb{P}(X \in [t, t+\Delta t], X \geq t)}{\mathbb{P}(X \geq t)} = \frac{\mathbb{P}(X \in [t, t+\Delta t])}{\mathbb{P}(X \geq t)}$$
$$\mathbb{P}(X \in [t, t+\Delta t] | X \geq t) = \mathbb{P}(X \in [0, t])$$
$$g(\Delta t) = \frac{g(t+\Delta t)}{g(t)} \Rightarrow g(\Delta t) g(t) = g(t+\Delta t)$$
$$g \sim \text{Exp}$$

We say that a random variable X is exponentially distributed to the parameter $\lambda > 0$ if

$$f_X(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

We write $X \sim \text{Exp}(\lambda)$ The cdf is given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

Poisson Distribution:

Ex: Arrival of customers at a store.

Let Y be the number of customers that arrive unitil time t.

$$1 - e^{-\lambda \Delta t} \approx \lambda \Delta t$$

Probability of a customer arriving is $\lambda \Delta t$ and probability of someone not arriving is $1 - \lambda \Delta t$. This represents a binomial distribution.

$$\Delta t = \frac{t}{n}$$
$$\text{Bin}\left(\frac{t}{\Delta t}, \lambda \Delta t\right) = \text{Bin}\left(\frac{t}{t/n}, \frac{\lambda t}{n}\right) = \text{Bin}\left(n, \frac{\lambda t}{n}\right)$$
$$\mathbb{P}(Y = k) = \lim_{n \to \infty} \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} = \lim_{n \to \infty} \frac{n!}{(n-k)!} \frac{(\lambda t)^k}{k!} \frac{1}{n^k} \left(1 - \frac{\lambda t}{n}\right)^{-k} \left(1 - \frac{\lambda t}{n}\right)^n$$

$$= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{(\lambda t)^k}{k!} \left(1-\frac{\lambda t}{n}\right)^{-k} \left(1-\frac{\lambda t}{n}\right)^n$$

$$= 1 \cdot \frac{(\lambda t)^k}{k!} \cdot 1 \cdot e^{-\lambda t}$$

$$\tilde{\lambda} := \lambda t$$

$$\mathbb{P}(Y=k) = \frac{\tilde{\lambda}^k}{k!} e^{-\tilde{\lambda}}$$

The discrete case of the exponential distribution is the Poisson distribution.
Let X be a discrete random variable. We say that X is Poisson distributed to the parameter $\lambda > 0$ if

$$\mathbb{P}(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

for all $k \in \{0,1,2,\ldots\}$.
We write $X \sim \text{Pois}(\lambda)$

1.2.3 Expectations

Ex: Roll a D6. Let X be the number of dots that show up. What is the average (or expected) number of dots?

$$\Omega = \{1,2,3,4,5,6\}$$

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = \frac{1}{6}$$

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5$$

So the avergae number of dots is 3.5. This is called the expected value of X, $\mathbb{E}[X]$
The expectation value of a discrete random variable X is defined as

$$\mathbb{E}[X] = \sum_{x_i} x_i \mathbb{P}(X=x_i) = \sum_{x_i} p_X(x_i)$$

For a continuous random variable it is

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

Now, what if we want to compute $\mathbb{E}[2X]$?
We get

$$\mathbb{E}[2X] = \sum_{x_i} 2x_i \mathbb{P}(X=x_i)$$

We can extend this to any function of X and get that

$$\mathbb{E}[g(X)] = \sum_{x_i} g(x_i) \mathbb{P}(X=x_i)$$

We can also compute the expected value for the various distributions that we have defined thus far.
$X \sim \text{Bernoulli}(p)$:

$$x_i \in \{0, 1\}$$
$$\mathbb{E}[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = 0 \cdot (1 - p) + 1 \cdot p = p$$

$X \sim U(\{k, \ldots, m\})$:

$$\mathbb{E}[X] = \sum_{i=k}^{m} i \frac{1}{m-k} = (m + k) \left(\frac{m-k}{2}\right) \left(\frac{1}{m-k}\right) = \frac{m+k}{2}$$

$X \sim \text{Bin}(n, p)$:

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$$
$$= \sum_{k=1}^{n} \frac{n!}{(n-k)!(k-1)!} p^k (1-p)^{n-k} = \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-1-(k-1)} np$$
$$= \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} np = np$$

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

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^{k-1} \lambda}{(k-1)!} e^{-\lambda} = \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = \lambda$$

$X \sim \text{Geom}(p)$:

$$\mathbb{E}[X] = 1 \cdot p + (1 - p)(1 + \mathbb{E}[X])$$
$$p\mathbb{E}[X] = p + (1 - p) = 1 \Rightarrow \mathbb{E}[X] = \frac{1}{p}$$

$X \sim U([a, b])$:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

$X \sim \text{Exp}(\lambda)$:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx$$
$$u = \lambda x, \; du = \lambda dx$$
$$\int_0^{\infty} x \lambda e^{-\lambda x} dx = -x e^{-\lambda x} \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx$$
$$= -\frac{1}{\lambda} e^{-\lambda x} \Big|_0^{\infty} = \frac{1}{\lambda}$$

We can also compute the expectations values of more than just X.
Ex: Compute $\mathbb{E}[X^2]$ for $X \sim U[a,b]$

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f_X(x)dx = \int_{\mathbb{R}} x^2 \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)dx = \frac{1}{b-a}\int_a^b x^2 dx = \frac{1}{b-a}\frac{x^3}{3}\Big|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

Ex2: Compute $\mathbb{E}[Y^2]$ for $X \sim \text{Exp}(\lambda)$

$$\mathbb{E}[Y^2] = \int_{\mathbb{R}} y^2 f_Y(y)dy = \int_{\mathbb{R}} y^2 \lambda e^{-\lambda y}\mathbb{1}\{y \geq 0\}dy$$
$$= \lambda \int_0^\infty y^2 e^{-\lambda y}dy = y^2 \frac{\lambda}{-\lambda}e^{-\lambda y}\Big|_0^\infty - \int_0^\infty 2y(-e^{-\lambda y})dy$$
$$= 0 + 2y\left(-\frac{1}{\lambda}\right)e^{-\lambda y}\Big|_0^\infty - \int_0^\infty 2\left(-\frac{1}{\lambda}\right)e^{-\lambda y}dy$$
$$= 0 - 2\left(-\frac{1}{\lambda}\right)\frac{1}{\lambda}e^{-\lambda y}\Big|_0^\infty = \frac{2}{\lambda^2}$$

Ex3: Compute $\mathbb{E}[e^y]$ for $Y \sim \text{Exp}(\lambda)$

$$\mathbb{E}[e^y] = \int_{\mathbb{R}} e^y f_Y(y)dy = \int_{\mathbb{R}} e^y \lambda e^{-\lambda y}\mathbb{1}\{y \geq 0\}dy = \lambda \int_0^\infty e^{-(\lambda-1)y}dy$$
$$= \left[\frac{\lambda}{-(\lambda-1)}e^{-(\lambda-1)}\right]_0^\infty = \frac{\lambda}{\lambda-1}$$

We can also deal with functions of multiple random variables.

$$\mathbb{E}[X+Y] = \sum_{x_i,y_i}(x_i+y_i)\mathbb{P}(X=x_i, Y=y_i)$$

y_i makes up all the values that Y can take so we can write

$$\sum_{y_i}\mathbb{P}(X=x_i, Y=y_i) = \mathbb{P}\left(\{X=x_i\} \cap \bigsqcup_{y_i}\{Y=y_i\}\right)$$
$$\bigsqcup_{y_i}\{Y=y_i\} = \Omega$$
$$\sum_{y_i}\mathbb{P}(X=x_i, Y=y_i) = \mathbb{P}(X=x_i)$$

Using this trick we can get

$$\mathbb{E}[X+Y] = \sum_{x_i}\sum_{y_i} x_i\mathbb{P}(X=x_i, Y=y_i) + \sum_{x_i}\sum_{y_i} y_i\mathbb{P}(X=x_i, Y=y_i)$$
$$= \sum_{x_i} x_i \sum_{y_i}\mathbb{P}(X=x_i, Y=y_i) + \sum_{y_i} y_i \sum_{x_i}\mathbb{P}(X=x_i, Y=y_i)$$
$$= \sum_{x_i} x_i\mathbb{P}(X=x_i) + \sum_{y_i} y_i\mathbb{P}(Y=y_i) = \mathbb{E}[X] + \mathbb{E}[Y]$$

We also saw earlier that $\mathbb{E}[2X] = 2\mathbb{E}[X]$. These two properties imply that $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ which is called the linearity of expectation.
Ex: Flip a coin 100 times. What is the expected number of streaks of 6 heads in a row?
$X :=$ number of streaks of 6.
$x_i :=$ streak of 6 happening with i.

$$\mathbb{E}[x_i] = 1 \cdot 12^6 + 0 \cdot \left(1 - \frac{1}{2^6}\right) = \frac{1}{64}$$
$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{95} x_i\right] = \sum_{i=1}^{95} \mathbb{E}[x_i] = \frac{95}{64}$$

Conditional Expectations:

$$\mathbb{E}[X|A] = \sum_{x_i} x_i \mathbb{P}(X = x_i|A)$$

Law of total probability for expectations:
Let $(A_i)_{i\in\{1,\dots,n\}}$ be a partition of Ω such that $\bigsqcup_{i=1}^{n} A_i = \Omega$ then

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i]\mathbb{P}(A_i)$$

Note that in general $\mathbb{E}[X^2]$ is not the same as $(\mathbb{E}[X])^2$.
Take the example of $f_X(x) = x\mathbb{1}_{[0,2]}(x)$

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)dx = \int_0^2 x^2 dx = \frac{8}{3}$$
$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f_X(x)dx = \int_0^2 x^3 dx = \frac{8}{4} = \frac{16}{4} = 4$$
$$(\mathbb{E}[X])^2 = \frac{64}{3} \neq 4 \Rightarrow (\mathbb{E}[X])^2 \neq \mathbb{E}[X^2]$$