# Math 180A

Introduction to Probability

Spring 2022
Taught by Professor Brett Kolesnik

# Table of Contents

# 1   Discrete Probability Distributions

In this section, we are mostly concerned with discrete sample spaces, or *finite* (or countably infinite) sample spaces.

## 1.1   Sample Space

Probability is the study of randomness, of uncertainty. Formally, we can think of this process as running a **random experiment**.

---
**Definition 1.1: Sample Space**

The **sample space** of an experiment is the set of all possible outcomes of that experiment. Namely, we say that
$$\Omega = \{\omega_1, \ldots, \omega_n\},$$
where each $\omega_i$ represents a possible outcome.

---

(Example.) When flipping a coin, we would have
$$\Omega_{\text{Coin}} = \{\text{Heads}, \text{Tails}\}.$$

The outcomes are assigned **masses** $m(\omega_i) \geq 0$ such that
$$m(\omega_1) + \cdots + m(\omega_n) = 1.$$

Here, we can think of the $m(\omega_i)$ as the **probability** that the outcome $\omega_i$ occurs.

(Example.) When rolling a regular die, we would have
$$\Omega_{\text{Die}} = \{1, 2, 3, 4, 5, 6\}.$$
Here, each $\omega \in \Omega_{\text{Die}}$ has a mass of $\frac{1}{6}$; that is,
$$\forall \omega \in \Omega_{\text{Die}}, m(\omega) = \frac{1}{6}.$$

## 1.2   Random Variables

We can use *random variables* to quantify the outcome.

---
**Definition 1.2: Random Variable**

Suppose we have an experiment whose outcome depends on chance. We can represent the outcome of the experiment by a captial Roman letter, such as $X$, called a **random variable** (RV).

We can think of a random variable $X$ as a function from the sample space $\Omega$ to the set of real numbers $\mathbb{R}$; that is,
$$X : \Omega \mapsto \mathbb{R}.$$

---

If the outcome $\omega \in \Omega$ has a random occurrence, then the value $X(\omega)$ will also be random.

(Example.) Suppose we wanted to flip a fair coin. It's obvious that the sample space $\Omega$ is just Heads

or Tails. So, we can define a random variable $X$ like

$$X = \begin{cases} 1 & \text{if Heads} \\ 0 & \text{if Tails} \end{cases}.$$

We can also define the random variable

$$X = \begin{cases} 21313 & \text{if Heads} \\ 0 & \text{if Tails} \end{cases}.$$

The point is that your random variable maps your outcomes (from your sample space) to numbers. In other words, you are quantifying your outcomes.

## 1.3   Events

Often, in probability, we are interested in the probability that a certain event will occur. For example, we might be interested in whether or not it will rain today, or whether or not we will win the lottery, or so on.

### Definition 1.3: Event

A subset of a sample space is an **event**.

The probability of an event $E$ is given by

$$\mathbb{P}(E) = \sum_{\omega \in E} m(\omega).$$

(Example.) If we roll a die, then $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event,

$$E = \text{``Roll an even number''},$$

is the subset $E = \{2, 4, 6\} \subset \Omega$.

## 1.4   Probability Distribution

If we have an outcome $\omega$, we can use the probability distribution function to get the probability that $\omega$ occurs.

### Definition 1.4: Probability Distribution

Let $\Omega$ be a discrete (finite or countable infinite) set. Then, the function

$$\mathbb{P} : \Omega \mapsto [0, 1]$$

is called a **probability distribution** on $\Omega$ if the following hold:

1. $\mathbb{P}(\omega) \geq 0$ for all $\omega \in \Omega$, and

2. $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$.

(Example.) If we roll a fair die, then we have

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

Then, it follows that

$$\mathbb{P}(i) = \frac{1}{6}$$

for all $i \in \Omega$.

## 1.5 Probability Mass Function

If $X$ is a random variable on $\Omega$, then we note that $\mathbb{P}(X = x)$ is a probability distribution on the set

$$\Omega_X = \{X(\omega) : \omega \in \Omega\}$$

of all possible values that $X$ can take.

> **Definition 1.5: Probability Mass Function**
>
> For a discrete random variable $X$, the function
>
> $$\mathbb{P} : \Omega_X \subset \mathbb{R} \mapsto [0,1]$$
>
> is called a **probability mass function** (PMF) of said random variable if the following hold:
>
> 1. $\mathbb{P}(x) \geq 0$ for all $x \in \Omega_X$, and
>
> 2. $\sum_{x \in \Omega_X} \mathbb{P}(x) = 1$.
>
> In other words, given a possible value of the random variable, the probability that the random variable takes that particular value is given by the function above.

We note that, for a random variable $X$,

$$\mathbb{P}(X = x) = \sum_{\omega : X(\omega) = x} m(\omega).$$

The $X = x$ inside the $\mathbb{P}(X = x)$ is shorthand notation for the **event**

$$\{\omega : X(\omega) = x\} \subset \Omega.$$

Additionally, instead of writing $\mathbb{P}(X = x)$, we may also write $p_X(x)$.

> (Example.) If we have a sample space $\Omega$ and a random variable $X : \Omega \mapsto \mathbb{R}$, then we can ask questions like "How likely is it that the value of $X$ is equal to 2?" This is the same as the probability of the event
>
> $$\{\omega : X(\omega) = 2\},$$
>
> or, equivalently,
>
> $$\mathbb{P}(X = 2)$$
>
> or
>
> $$p_X(2).$$

> (Example.) Suppose again we have a fair die. If $X$ is a random variable that takes the value 1 if the roll is 1 or 2, and the value 0 otherwise, then
>
> $$\mathbb{P}(X = 1) = \frac{1}{3}$$
>
> and
>
> $$\mathbb{P}(X = 0) = \frac{2}{3}$$

is the probability distribution of $X$ on the set $\Omega_X = \{0, 1\}$.

So, effectively, we can think of a probability mass function as a probability distribution function for values that the random variable can take. **We will be using this a lot.**

## 1.6 Cumulative Distribution Function

While we have the probability mass function, we also have the cumulative distribution function.

> **Definition 1.6: Cumulative Distribution Function**
>
> The **cumulative distribution function** (CDF) of a discrete random variable $X$ is the function given by
> $$F_X(x) = \mathbb{P}(X \le x) = \sum_{t:t \le x} \mathbb{P}(X = t).$$

So, whereas the PMF gives us the probability of a random value taking a specific value, the CDF gives us the probability that a random variable takes on a value *less than or equal to* a specific value.

---

(Example.) Suppose we're rolling a fair die, and suppose we have the random variable $X$ that takes the value 1 if $\omega \in \{1\}$ and the value 0 if $\omega \in \{2, 3, 4, 5, 6\}$. Then, $X = 1$ if we roll a 1 and $X = 0$ otherwise. Then,
$$\mathbb{P}(X = 1) = m(1) = \frac{1}{6} \text{ and } \mathbb{P}(X = 0) = m(2) + \cdots + m(6) = \frac{5}{6}.$$

---

(Example.) Suppose we roll a die 3 times in a row. Let $X$ be the number of 1's that we roll. Then,
$$X = X_1 + X_2 + X_3$$
where each $X_i$ has the same distribution, i.e. $\mathbb{P}(X_i = 1) = \frac{1}{6}$ and $\mathbb{P}(X_i = 0) = \frac{5}{6}$ for each $i \in [1, 3]$.

---

(Example.) Again, suppose we roll a die 3 times in a row. Then,
$$X = X_1 + X_2 + X_3.$$

Suppose we wanted to find the probability that at most *2* of those rolls are 1's. We essentially need to calculate
$$\mathbb{P}(X \le 2) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2).$$

There are three possibilities:

(a) $\mathbb{P}(X = 0)$: All 3 rolls are not 1's. Here, the probability that we don't get a 1 is $\frac{5}{6}$. So, the probability that all three rolls are not 1's is given by $(5/6)^3$. There is only $\binom{3}{0} = 1$ way we can possibly obtain zero 1's. To visualize this, let $x$ be some non-one integer. Then,

       x x x.

(b) $\mathbb{P}(X = 1)$: Exactly 1 of the 3 rolls is a 1. The probability that we get a 1 is $\frac{1}{6}$ and the probability that we don't get a 1 is $\frac{5}{6}$. We note that there are $\binom{3}{1} = 3$ possible ways we can obtain one 1. To visualize this, let $x$ be some non-one integer. Then,

       1 x x
       x 1 x
       x x 1.

(c) $\mathbb{P}(X = 2)$: Exactly 2 of the 3 rolls is a 1. We note that there are $\binom{3}{2}$ ways we can possibly get two 1's. To visualize this, let $x$ be some non-one integer. Then,

```
1 1 x
x 1 1
1 x 1.
```

Putting this together, we have:

$$\mathbb{P}(X \leq 2) = \binom{3}{0}\left(\frac{5}{6}\right)^3 + \binom{3}{1}\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2 + \binom{3}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right) \approx 99.54\%.$$

## 1.7   Review of Set Theory

Recall that events are subsets. In particular, let $A, B \subset \Omega$ be two events. Then:

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

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

- Difference: $A \setminus B = \{\omega \mid \omega \in A \text{ and } \omega \notin B\}$.

- Complement: $A^C = \Omega \setminus A$.

Two events $A$ and $B$ are said to be **disjoint** if $A \cap B = \emptyset$. If two events are disjoint, then it is impossible for them to both occur at the same time.

## 1.8   Properties of Probability Distribution

We now introduce our first theorem of this class.

> **Theorem 1.1**
>
> Suppose that $\mathbb{P}$ is a probability distribution on a discrete set $\Omega$. Then,
>
> 1. $\mathbb{P}(E) \geq 0$ for all events $E \subset \Omega$.
>
> 2. $\mathbb{P}(\Omega) = 1$.
>
> 3. If $E \subset F \subset \Omega$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$.
>
> 4. If $A \cap B = \emptyset$ are disjoint, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
>
> 5. $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$ for all events $A \subset \Omega$.

*Proof.* We'll prove each of the following statements in this theorem.

1. We know that $E$ is a subset of $\Omega$. Take some $\omega \in E$. Then, we know that its mass, $m(\omega) \geq 0$. Thus, it follows that
$$\sum_{\omega \in E} m(\omega) \geq 0.$$

2. Recall that $\Omega$ is the set of all possible outcomes. Therefore, it follows that $\mathbb{P}(\Omega) = 1$; if this is false, then this implies that there is at least one outcome that isn't in $\Omega$.

3. Using (1) as a baseline, suppose that $\mathbb{P}(F) = f$. Since $E$ can only have elements from $F$, it follows that $\mathbb{P}(E) \leq f$; if this statement is false, this implies that $E$ has elements that aren't in $F$, which cannot be the case.

4. Since $A \cap B = \emptyset$, it follows that

$$\mathbb{P}(A \cap B) = \sum_{\omega \in A \cap B} \mathbb{P}(\omega) = \sum_{\omega \in A} \mathbb{P}(\omega) + \sum_{\omega \in B} \mathbb{P}(\omega) = \mathbb{P}(A) + \mathbb{P}(B).$$

The key here is that we are not double-counting anything.

5. Recall that $A \cup A^C = \Omega$ and $A \cap A^C = \emptyset$. In particular, since $A \cap A^C = \emptyset$, then

$$\mathbb{P}(A \cup A^C) = \mathbb{P}(A) + \mathbb{P}(A^C).$$

But, since $A \cup A^C = \Omega$ and $\mathbb{P}(\Omega) = 1$, we know that

$$1 = \mathbb{P}(A) + \mathbb{P}(A^C).$$

Therefore, it follows that

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

This concludes the proof. □

Looking at #4 in the previous theorem, we can actually generalize this.

---

**Theorem 1.2**

If $A_1, \ldots, A_n$ are pairwise disjoint (i.e. $\bigcap_{i \in [1,n] \subset \mathbb{Z}} A_i = \emptyset$), then

$$P\left(\bigcup_{n=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

---

*Proof.* One way we can go about this is to take advantage of the fact that $A_1, \ldots, A_n$ are pairwise disjoint. We will use induction on $n$.

- <u>Base Case:</u> Suppose $n = 1$. Trivially, $A_1$ is pairwise disjoint since it's the only set and so $\mathbb{P}(A_1) = \mathbb{P}(A_1)$. Likewise, $n = 2$ is satisfied by the previous theorem.

- <u>Inductive Step:</u> Suppose that this holds for $n$. We now want to show that this holds for $n + 1$. To do so, we note that
$$A_1 \cap \cdots \cap A_n = \emptyset$$
and
$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n).$$

So, we can define $A = A_1 \cup \cdots \cup A_n$. We now introduce the set $A_{n+1}$; suppose that $A_{n+1} \cap A = \emptyset$. Then, it follows that

$$\mathbb{P}(A \cup A_{n+1}) = \mathbb{P}(A) + \mathbb{P}(A_{n+1}) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) + \mathbb{P}(A_{n+1})$$

This concludes the proof. □

**Remarks:**

- The following consequence of the previous theorem is an extremely useful tool for calculating the probability of events.

- Often, it is difficult to find $\mathbb{P}(E)$ directly, and it is easier to "split the job up" into doable subtasks.

## 1.9   Law of Total Probability

This brings us to the *Law of Total Probability*.

> ### Theorem 1.3: Law of Total Probability (LoTP)
>
> Let $E \subset \Omega$ be an event, and let $A_1, \ldots, A_n$ be a partition of $\Omega$ (that is, a pairwise disjoint collection of sets that "cover" the sample space $\bigcup_{i=1}^n A_i = \Omega$). Then, we have that
> $$P(E) = \sum_{i=1}^n P(E \cap A_i)$$

> *Proof.* Note that $E$ is the pairwise disjoint union of the sets $E \cap A_1, \ldots, E \cap A_n$. Thus, we can just apply the previous theorem.  $\square$

**Remark:** While it might be difficult to find $P(E)$ directly, if you pick the $A_i$'s wisely, it can become easy to find each of the $P(E \cap A_i)$'s.

> ### Corollary 1.1
>
> For any two events $A$ and $B$,
> $$P(A) = P(A \cap B) + P(A \cap B^C)$$

**Remark:** This holds since $B$, $B^C$ is a parition of $\Omega$.

> ### Theorem 1.4: I
>
> $A$ and $B$ are subsets of $\Omega$, then
> $$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

> *Proof.* Recall that
> $$P(A \cup B) = \sum_{\omega \in A \cup B} m(\omega).$$
>
> Now, if $\omega$ is in exactly one of the two sets, then it is only counted once (hence the first and second term on the right-hand side). However, if $\omega$ is in both $A$ and $B$, then we would be double-counting since it's counted for in $P(A)$ and $P(B)$. So, we need to subtract it (hence, the last term on the right-hand side).  $\square$

**Remark:** If $A \cap B = \emptyset$, then $P(A \cap B) = 0$.

## 1.10   Example Probability Distributions

We now talk about two types of distributions: uniform and geometric distributions.

### 1.10.1    Uniform Distribution

> **Definition 1.7: Uniform Distribution**
>
> The *uniform distribution* on a <u>finite</u> sample space $\Omega$ containing $n$ elements is the function $m$ defined by
> $$m(\omega) = \frac{1}{n}$$
> for every outcome $\omega \in \Omega$.

For example, when flipping a fair coin, there is only two possiblities: heads or tails. So,

$$m(\text{Heads}) = \frac{1}{2}.$$

A nice property of the uniform distribution is that, for all events $E \subset \Omega$, we simply have that

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

### 1.10.2    Infinite Sample Sizes & Geometric Distribution

The same definitions that we discussed earlier apply the same way in the infinite case. However, notice that the rule

$$\sum_{i=1}^{\infty} \mathbb{P}(\omega_i) = 1$$

means that this infinite series *converges*, and it converges to 1 (it is not just a usual sum). Now, when $\Omega$ is countably infinite, we further assume, in the definition of probability distribution, that

$$\mathbb{P}\left(\bigcup_{i \in I} E_i\right) = \sum_{i \in I} \mathbb{P}(E_i)$$

for all (possibly countably infinite) collections of pairwise disjoint sets $\{E_i \mid i \in I\}$.

Now that we know the basics of infinite sample size, we can now consider the **geomtric distribution**[1]

$$P(X = k) = p(1 - p)^{k-1}$$

for $k = 1, 2, \ldots$ and $P(X = x) = 0$ for all other $x$.

> (Example: Geometric Distribution.) To see this, suppose a coin flips "Tails" with probability $p$. Then, the random variable, $X$ is the number of flips until we flip "Tails" for the first time, has this distribution. For example, if we flip "Heads" twice and get "Tails" on the third attempt, then $X = 3$. Indeed, for this to happen on flip $k$, we need all of the previous $k - 1$ flips to be "Heads," and then the next flip to be "Tails."
>
> Thus, the probability that we only get "Tails" on the third attempt (i.e., "Heads" on the first two attempts) is given by
> $$\mathbb{P}(X = 3) = (1 - p)(1 - p)p = (1 - p)^2 p = (1 - p)^{3-1}p.$$

To check that this is a bonafide probability distribution, we note that

$$\sum_{k=1}^{\infty} P(X = k)$$

---

[1]This will be discussed in detail later on.

is equal to

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

Note that the second-to-last step is from the geometric series

$$\sum_{i=0}^{\infty} \alpha^i = \frac{1}{1-\alpha}$$

for $|\alpha| < 1$.

# 2 Continuous Probability Distributions

Now, we will discuss the case where there is a continuum of possible values that a RV can take. For example, rather than discrete choices like 1, 2, 3, 4, 5, we will instead he dealing with things like the time until the first customer appears at a store, or the lifetime of a lightbulb.

## 2.1 Probability Density Function

Recall that, in the discrete case, a random variable's probability mass function (PMF)

$$p_X(x) = \mathbb{P}(X = x)$$

has the property that

$$\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x).$$

We now define the analog to the PMF. In particular, we want to find a probability density function $f$ such that

$$\mathbb{P}(X \in A) = \int_A f(x)dx,$$

where $A \subseteq \mathbb{R}$ is some arbitrary region. Note that $f(x)$ is not a probability.

---

**Definition 2.1: Probability Density Function**

Let $X$ be a continuous, $\mathbb{R}$-valued random variable. A **probability density function** (PDF) for $X$ is a $\mathbb{R}$-valued, non-negative function $f$ that satisfies

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx$$

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

---

We note that

$$\mathbb{P}(X = x) = \int_x^x f(x)dx = 0$$

for any $x$. This means that

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b).$$

For example, a uniform random variable on an interval $I \subset \mathbb{R}$ has PDF

$$\boxed{f = \frac{1}{\text{length}(I)}}.$$

## 2.2 Cumulative Distribution Function

While we have the probability density function, we also have the cumulative distribution function.

---

**Definition 2.2: Cumulative Distribution Function**

The **cumulative distribution function** (CDF) of a continuous random variable $X$ is the function given by

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(t)dt.$$

---

## 2.3    Relationship Between PDF and CDF

We can relate the PDF and the CDF by the following theorem:

> **Theorem 2.1**
>
> Let $X$ have CDF $F_X$ and PDF $f_X$. Then, $F'(x) = f(x)$.

*Proof.* Note that $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$. Hence, by the Fundamental Theorem of Calculus, it follows that

$$F'_X(x) = f_X(x),$$

as desired.                                                                                        □

> **Important Note 2.1**
>
> The PDF is the derivative of the CDF.

---

(Example.) Suppose we have a dart board with unit radius. Suppose we throw a dart at the target. The sample space is the unit disk

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

The unit circle has area $\pi(1)^2 = \pi$. Supposing a dart lands at uniformly random position on the target, we would have the PDF

$$f(x, y) = \frac{1}{\pi}$$

for a random throw $(X, Y)$. Note that, because this is two-dimensional, we consider the *area* as opposed to the length of the interval.

To find the probability of landing in a certain region, we would have to integrate over that region; more specifically, we have to integrate that uniform density $\frac{1}{\pi}$ by that region. For instance, if the "bullseye" region of the target is the center circle $B$ of radius $\frac{1}{5}$, then the probability of getting a "bullseye" would be

$$\frac{\text{area}(B)}{\text{area}(D)} = \frac{\pi(1/5)^2}{\pi} = \frac{1}{25}.$$

So, we should expect approximately 1 in every 25 throws to be a "bullseye."

Now, let $D$ be the distance from the center to the point $(X, Y)$ where a uniformly thrown dart lands. We note that

$$D = \sqrt{X^2 + Y^2} \in [0, 1].$$

What is the distribution of this random variable? We should not expect $D$ to have a uniform distribution. For instance, notice that there are more points at distance $\geq 1/2$ from the center than there are points at distance $\leq 1/2$ from the center. So, we expect

$$\mathbb{P}(D \in [0, 1/2]) < \mathbb{P}(D \in [1/2, 1])$$

although both of these sub-intervals of $[0, 1]$ have the same length. Recall that $f(d) = F'(d)$. Then,

$$F(d) = \mathbb{P}(D \leq d) = \mathbb{P}(X^2 + Y^2 \leq d^2) = \frac{\pi d^2}{\pi} = d^2.$$

Notice here that $\pi d^2$ is the area of the *inner* circle. Therefore, $f(d) = 2d$.

---

(Example Problem.) Let $U$ be a uniform random variable on $[0, 1]$, and consider the random variable

$$X = U^2.$$

Find the PDF of $X$.

We know that $X$ has PDF
$$F_X(x) = \mathbb{P}(X \le x).$$

We also know that $U$ is a uniform RV on $[0, 1]$, so its PDF is given by

$$f_U(u) = \begin{cases} \frac{1}{1-0} = 1 & \text{if } u \in [0, 1] \\ 0 & \text{otherwise} \end{cases}.$$

So, we have that

$$\begin{aligned} \mathbb{P}(X \le x) &= \mathbb{P}(U^2 \le x) \\ &= \mathbb{P}(U \le \sqrt{x}) \\ &= \int_{-\infty}^{\sqrt{x}} f_U(t)dt \\ &= \int_0^{\sqrt{x}} 1 dt \\ &= \sqrt{x}. \end{aligned}$$

This tells us that the CDF is
$$F_X(x) = \sqrt{x}.$$

Then, to find the PDF, we can just take the derivative of the CDF, like so:

$$\frac{d}{dx}F_X(x) = \frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}.$$

## 2.4    Exponential Random Variable

This random variable is useful when studying events occurring at random times. For example, lifetime of a lightbulb, time until the next customer, time until the next earthquake, etc.

Recall that $F(x) = \mathbb{P}(X \le x)$. Hence,
$$1 - F(x) = \mathbb{P}(X > x).$$

This is sometimes denoted by $S(x) = 1 - F(x)$ and is referred to as the **survival function** of the random variable $X$. Note that if $X$ is a random time, e.g. the lifetime of a lightbulb, then $S(x)$ is the probability of "surviving" until time $x$.

A very special type of continuous random variable is the exponential random variable with rate $\lambda > 0$. This is the random variable with survival function

$$S(x) = e^{-\lambda x}.$$

That is, the probability of survival until time $X$ decays exponentially with rate $\lambda$. Note that $F(x) = 1 - e^{-\lambda x}$, and so $f(x) = \lambda e^{-\lambda x}$ for $x \ge 0$ and $f(x) = 0$ otherwise is its PDF.

### 2.4.1    Memoryless Property

The exponential RV is very important because it is the only continuous RV with a special property, known as the **memoryless property**. Then, given that an exponential RV has survived until time $x$, the probability

that it survives for $y$ amount of time longer (i.e. until time $x + y$) is the same as the probability of just surviving until time $y$.

For example, if an exponential lightbulb has survived until time $x$, then given this, the probability of surviving until time $x + y$ is the same as the probability of a brand new lightbulb.

# 3    Combinatorics

We will be studying various counting techniques. In probability, we are often interested in the probability of a certain event. Finding such a probability usually involves considering/counting all of the different ways in which it could possibly occur.

## 3.1    Basic Principle of Counting

Suppose that an experiment involves $r$ independent stages, i.e. stages that have no effect on each other. Suppose that, in each stage $1 \le i \le r$, there are $n_i$ possible outcomes. Then, the total number of possible outcomes of the full experiment is the product $n_1 n_2 \ldots n_r$.

---

(Example.) A menu consists of 2 appetizers, 5 main dishes, and 2 desserts. Then, there are $2(5)(2) = 20$ possible meals to choose from. In particular:

- There are 2 appetizers that we can pick.

- For each appetizers, there are 5 dishes that we can pick.

- For each appetizers, there are 2 desserts we can pick.

---

(Example: Birthday Problem.) How many people do we need to have in a room so that it is likely (e.g. more than 50 percent) that two people will have the same birthday?

Assume that everyone is equally likely to be born on any one of the 365 days in a year. We note that the uniform probability assumption is not so realistic (less likely to be born on weekends, some months more likely than others, etc.). However, our calculation gives a good *upper bound* on the number of people needed in the room, since if the probabilities are non-uniform then the probability of having two people with the same birthday increases.

Suppose that there are $r$ people in the room. Then, there are $(365)^r$ possible birthdays: there are 365 choices for the first person, 365 choices for the second person, ..., and 365 choices for the $r$th person. *However*, there are only $(365)(364)\ldots(365 - r + 1)$ possible ways that they could all have different birthdays. In particular:

- There are 365 options for the first person.

- THere are 364 options for the second person.

- ...

- There are $365 - (r - 1)$ options for the $r$th person.

The probability that two people will have the same birthday when there are $r$ people in the room is

$$1 - \frac{(365)_r}{(365)^r}.$$

The probability that no one will have the same birthday is $\frac{(365)_r}{(365)^r}$ because $(365)^r$ is the total number of possibilities for all the birthdays, and $(365)_r$ is the total number of possibilities where everyone has different birthdays. Then, by the complement rule, the above value is the probability that at least two people with the same birthday. By plotting this function, we have that $r > 23$.

This is sometimes referred to as the **Birthday Paradox**. Despite this, it is not a paradox. This is because the total number of *pairs* of people is given by

$$\frac{23(22)}{2} = 253.$$

To see why this is the case, we have:

- 23 ways to pick the first person for the pair.

- 22 ways to pick the second person for the pairs.

We then divide by 2 because a pair $(A, B)$ is the same thing as $(B, A)$. That being said, this is much more comparable to 365, and all we need is for *one* of these pairs to be a match.

---

**Definition 3.1**

Let $0 \leq r < n$ be integers. Then,

$$(n)_r = n(n-1)\ldots(n-r-1)$$

is known as the **falling factorial**.

---

**Definition 3.2**

We denote $(n)_n = n(n-1)\ldots1$ as $n!$, and $0! = 1$, and denote this as $n$ **factorial**.

---

We note that $n!$ is the total number of possible ways to permute (order) a list of $n$ distinguishable objects.

---

(Example.) There are $3! = 3(2)(1)$ possible ways to arrange three people in a line.

- There are 3 ways to pick the first person to be at the front of the line.

- There are 2 ways to pick the second person to be next in line.

- Finally, there is only 1 way to pick the one line person to be last in line.

---

## 3.2    Permutations

**Definition 3.3**

A **permutation** of a finite set $A$ is a bijective mapping from $A$ to $A$.

---

**Remarks:**

- We often, but not always, use the Greek letters $\pi$ or $\sigma$ to denote permutations.

- Note that any set $A$ of size $|A| = n$ is in bijective correspondence with $[n] = \{1, 2, \ldots, n\}$. That is, we can enumerate $A = \{a_1, \ldots, a_n\}$. So, we will usually only discuss permutations of $[n]$.

---

(Example.) Consider the permutation of $\sigma$ of $[4] = \{1, 2, 3, 4\}$ such that

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}.$$

Here,

$$\sigma(1) = 2$$

$$\sigma(2) = 1$$
$$\sigma(3) = 4$$
$$\sigma(4) = 3.$$

In particular, the top row is a list of elements, and the bottom is the **rearrangement** of them given by $\sigma$. So, in this example, 2 becomes the first element, 1 becomes the second element, and so on.

Occasionally, we might just write $\sigma = 2143$.

### Definition 3.4

Let $S_n$ denote the set of all permutations of $[n]$, sometimes known as the **symmetric group** of degree $n$.

**Remark:** $|S_n| = n!$.

(Example.) The 6 permutations of $[3]$ are:

- 123

- 132

- 213

- 231

- 312

- 321

### Theorem 3.1: Stirling's Approximation

As $n \mapsto \infty$, we have

$$\frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} \mapsto 1.$$

**Remark:** Hence, for large $n$, we have $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$.

Recall that a **fixed point** $x$ of a function $f$ is a point for which $f(x) = x$. Suppose we select a uniformly random permutation of $[n]$. What is the probability $p_k(n)$ that it will have exactly $k$ fixed points? Put it differently, if we put $n$ books on a shelf randomly, what is the probability that $k$ of them will happen to end up in their proper place on the shelf? In a future lecture, we will see that $p_0(n) \approx \frac{1}{e}$.

### Definition 3.5

Let $\sigma$ be a permutation of $[n]$. We call $i \in [n]$ a **record** if $\sigma(i) > \sigma(j)$ for all $j < i$.

**Remarks:**

- Informally, $i$ is a record if $\sigma(i)$ is larger than all of the previous values of $\sigma$.

- Trivially, $i = 1$ is a record.

## 3.3    Combinations

Suppose we have a set $A$ of size $|A| = n$. How many subsets of $A$ are there? How many of these are of size $k$?

---
**Definition 3.6**

The number of ways to choose $k$ elements from a set of $n$ distinguishable objects is denoted by $\binom{n}{k}$.

---

**Remarks:**

- If $|A| = n$, then there are $\binom{n}{k}$ subsets of $S \subset A$ of size $|S| = k$, since each subset corresponds to a way of choosing $k$ elements from the set of $n$ elements.

- The number $\binom{n}{k}$ is also known as the **binomial coefficient**.

---
**Theorem 3.2**

For $0 \leq k \leq n$, we have
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

---

**Remarks:**

- Recall that $0! = 1$.

- Recall that $\frac{n!}{(n-k)!} = (n)_k$.

- So, we can say that $\binom{n}{k} = \frac{(n)_k}{k!}$.

One thing to note is that the relationship

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

gives us a recursive method of computing binomial coefficients. This is known as the famous **Pascal's Triangle**. Another thing to notice is that

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

and so it follows that

$$\binom{n}{k} = \binom{n}{n-k}.$$

---

(Example.) Why does a Four of a Kind beat a Full House in Poker?

- Recall that there are 52 cards in a deck. In Poker, you get 5 cards.

- A Four of a Kind means that you get 4 of one of the kind of cards and one other card (e.g. all 4 aces and a 2).

- A Full House is when you get 3 of one kind and 2 of the other (e.g. 3 aces and 2 3's.).

---

How many ways can we get a Full House?

- We need to get two types of cards. There are $13 \cdot 12$ ways to do this.

- Now, we need to make the full house. For one of the cards, there are 4 different suits, of which we want 3 suits. Thus, $\binom{4}{3}$.

- For the other card, there are again 4 different suits, of which we want 2 suits. Thus, $\binom{4}{2}$.

This gives us

$$13 \cdot 12 \cdot \binom{4}{3}\binom{4}{2}.$$

How many ways can we get a Four of a Kind?

- There are 13 ways to choose the first card (that we need 4 of one kind for).

- There are now 48 cards left. We just need to pick one card to be our extra card.

This gives us

$$13 \cdot 48.$$

The probability of a Full House is given by

$$\frac{13 \cdot 12 \cdot \binom{4}{3}\binom{4}{2}}{\binom{52}{5}} \approx 0.14\%.$$

The probability of a Four of a Kind is given by

$$\frac{13 \cdot 48}{\binom{52}{5}} \approx 0.024\%.$$

For both of these cases, the $\binom{52}{5}$ came from the fact that we get 5 cards from a deck of 52.

---

(Example Problem.) Compute (with explanation) the probability that a poker hand contains (exactly) one pair (*aabcd* with *a*, *b*, *c*, *d* distinct face values). (*Answer:* $\approx 42.3\%$.)

First, there are 13 ways we can pick one kind of card. There are $\binom{4}{2}$ ways to pick 2 of the same card under different suits. Thus, the number of ways we can pick a pair of one kind of card is given by

$$13\binom{4}{2}.$$

There are now $\binom{12}{3}$ ways to pick the 3 remaining kinds of cards, and for each card there are $\binom{4}{1}$ ways to pick one card with some suit. Thus, this gives us

$$\binom{12}{3}\left(\binom{4}{1}\right)^3.$$

Combining this, we have

$$\frac{13\binom{4}{2}\binom{12}{3}\left(\binom{4}{1}\right)^3}{\binom{52}{5}} \approx 42.3\%,$$

as desired.

(Example Problem.) A researcher requires an estimate for the number of trout in a lake. To this end, she captures 50 trout, marks each fish, and releases them into the lake. Two days later she returns to the lake and captures 80 trout, of which 16 are marked. Suppose that the lake contains $n$ trout. Find the probability $L(n)$ that 16 trout are marked in a sample of 80.

We know that:

- There are $n$ trouts in the lake.

- 50 of the trouts in the lake are marked.

- We caught 80 trouts, of which 16 of them are marked.

I claim that $L(n)$ is given by the formula

$$L(n) = \frac{\binom{50}{16} \cdot \binom{n-50}{64}}{\binom{n}{80}}.$$

To see why this is the case, we have:

- There are $\binom{n}{80}$ ways to catch 80 trouts from the $n$ total trouts in the lake.

- We know that 16 trouts that we catch have to be marked. We also know that there are 50 marked trouts in the lake in total. Thus, there are $\binom{50}{16}$ ways to get 16 marked trouts from the 50 marked trouts in the lake.

- Since we know that the 16 trouts that we caught are marked, it follows that $80 - 16 = 64$ trouts that we caught are not marked. We also know that there are $n - 50$ trouts in the lake that aren't marked. Thus, there are $\binom{n-50}{64}$ ways to catch 64 unmarked trouts from the $n-50$ unmarked trouts in the lake.

## 3.4   Binomial Distribution

There is an important probability distribution related to the binomial coefficients, called the **Binomial Distribution**. But, we first need to describe the concept of a **Bernoulli trial**.

### Definition 3.7: Bernoulli Trial

A **Bernoulli trial** is a simple experiment that is either a *success* or *failure*. More specifically, it is a discrete random variable that either takes the value 1 (success) or 0 (failure).

Moreover, a Bernoulli($p$) trial is one in which the probability of success is $p$. Hence, its PMF is

$$\mathbb{P}(X = 1) = p$$

and

$$\mathbb{P}(X = 0) = 1 - p.$$

For example, flipping "Tails" when tossing a fair coin is a Bernoulli(1/2) trial.

We haven't defined what independence means in probability, but informally, a series of events $E_1, \ldots, E_n$ are independent if their outcomes "do not affect" each other. For example, when we flip a coin, whatever we get on the first flip won't affect what we get on the second flip. So, in this case, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right) = \prod_{i=1}^{n} \mathbb{P}(E_i).$$

In other words, the probability that they all occur is just the product of the individual probabilities.

(Example Problem.) A magician gives you a coin that comes up "Heads" with some probability $p \in (0,1)$. This probability $p$ is unknown, and perhaps you even have reason (e.g., the magician wears a suspicious looking grin, etc.) to suspect that the coin is biased, $p \neq 1/2$.

Show that this coin can be "turned into" a fair coin, by redefining what "Heads" means, in the following way: Flip the coin twice in a row. If the first flip is "Heads" and the second flip is "Tails", call the result "Heads New". On the other hand, if the first flip is "Tails" and the second flip is "Heads", call the result "Tails New". Otherwise, flip the coin twice in a row again, and keep repeating this procedure until either "Heads New" or "Tails New" has been determined.

Show that $\mathbb{P}(\text{"Heads New"}) = 1/2$.

Let $\mathbb{P}(\text{HT})$ be the probability that we get a heads followed by a tails. We know that the probability of getting a heads is $p$, and the probability of getting a tails is $1 - p$. So, the probability that we get a heads followed by a tails in our first try is given by

$$p(1-p).$$

If we don't get a heads and then a tails, then we either got a heads and heads or tails and tails (otherwise, we would lose if we got tails and heads). The probability of getting a heads and heads or tails and tails is given by
$$p^2 + (1-p)^2.$$

So, the probability that we initially get a heads and heads or tails and tails is given by

$$(p^2 + (1-p)^2)\mathbb{P}(\text{HT}),$$

where $\mathbb{P}(\text{HT})$ is due to us trying again. Therefore, we have

$$
\begin{aligned}
\mathbb{P}(\text{HT}) &= p(1-p) + (p^2 + (1-p)^2)\mathbb{P}(\text{HT}) \\
\implies \mathbb{P}(\text{HT}) &= p(1-p) + \mathbb{P}(\text{HT})p^2 + \mathbb{P}(\text{HT})(1-p)^2 \\
\implies \mathbb{P}(\text{HT}) - \mathbb{P}(\text{HT})p^2 - \mathbb{P}(\text{HT})(1-p)^2 &= p(1-p) \\
\implies \mathbb{P}(\text{HT})(1 - p^2 - (1-p)^2) &= p(1-p) \\
\implies \mathbb{P}(\text{HT})(1 - p^2 - (1 - 2p + p^2)) &= p(1-p) \\
\implies \mathbb{P}(\text{HT})(1 - p^2 - 1 + 2p - p^2) &= p(1-p) \\
\implies \mathbb{P}(\text{HT}) &= \frac{p(1-p)}{1 - p^2 - 1 + 2p - p^2} \\
\implies \mathbb{P}(\text{HT}) &= \frac{p(1-p)}{2p - 2p^2} \\
\implies \mathbb{P}(\text{HT}) &= \frac{\cancel{p(1-p)}}{2\cancel{p(1-p)}} \\
\implies \boxed{\mathbb{P}(\text{HT}) = \frac{1}{2}}.
\end{aligned}
$$

## Definition 3.8: Binomial Distribution

Let $n \geq 1$ be an integer and $P \in [0,1]$. Let $N$ be the number of "successes" in a series of $n$ independent Bernoulli($p$) trials. Then, we say that $N$ has the Binomial($n, p$) distribution.

Its PMF[2] is given by

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

for $0 \leq k \leq n$ and $\mathbb{P}(N = x) = 0$ otherwise. To see why this is the case,

- There are $\binom{n}{k}$ ways to choose which $k$ of the $n$ trials will be successful. Each of these $k$ trials will be a success with probability $p$, and each of the remaining $n - k$ trials will be a failure with probability $1 - p$.

- Since these trials are independent, we can multiply everything together: each of the $\binom{n}{k}$ outcomes that would cause $N - k$ to have probability $p^k (1-p)^{n-k}$.

How do we see that this is a legitimate probability distribution? In other words, how do we know that this all adds up to 1?

**Theorem 3.3: Binomial Theorem**

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

With this in mind, we have

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

---

[2]Probability Mass Function

# 4 Conditional Probability

Let $A$ be an event. Recall that $\mathbb{P}(A)$ is the probability that $A$ occurs. Now, suppose that we're given some additional information. Then, this information may not completely determine whether $A$ has occurred, but it might give us some valuable partial information.

> (Example.) Suppose that a magician rolls a fair die. Let $A$ be the event that we roll a 6. Then, we know that
> $$\mathbb{P}(A) = \frac{1}{6}.$$
> Now, let's suppose that the magician tells you that the result is an even number (either 2, 4, or 6), but does not yet reveal the full result of the roll to you. How does the probability of $A$ change? It's certainly not $\frac{1}{6}$ since we know that it has to be an even number. Intuitively, the answer is $\frac{1}{3}$; this is particularly because the die originally was uniform, so there is no reason why – when the magician got an even number – the even numbers have a heavier weight.

## 4.1 What is Conditional Probaiblity?

Conditional probability is the study of probability under the presence of partial information.

> **Definition 4.1: Conditional Probability**
>
> Let $A$ and $B$ be two events such that
>
> - $A$ is the event of interest.
>
> - $B$ is the event that encodes the partial information that we have.
>
> Suppose that $\mathbb{P}(B) > 0$. Then, the **conditional probability** of $A$, given that $B$ has occurred, is denoted by $\mathbb{P}(A|B)$.

**Remarks:**

- We require $\mathbb{P}(B) > 0$; if $\mathbb{P}(B) = 0$, then this would imply that $B$ would never occur anyways.

- All of the given information is on the *right* of the bar.

In our example above, $A$ would be the event of interest and $B$ is the event coming from the magician telling us that the roll was an even number.

## 4.2 Finding Conditional Probability

How do we find $\mathbb{P}(A|B)$?

> (Example, Continued.) In the previous example, it's clear that the probability of rolling a 6 should change from $\frac{1}{6}$ to $\frac{1}{3}$ once we found out that the roll is an even number. Additionally, the roll is uniformly random – now, we know that there is one of *three* possible numbers, so it should still remain uniformly random, but just on the sample space $\{2, 4, 6\}$ instead of the original sample space $\{1, 2, 3, 4, 5, 6\}$.
>
> So, if $A$ is the event that we rolled a 6, and $B$ is the even that we rolled an even number, notice that:
>
> - $\mathbb{P}(A) = \frac{1}{6}$: This is the probability that we roll a 6.
>
> - $\mathbb{P}(B) = \frac{1}{2}$: This is the probability that we will any even number.
>
> - $\mathbb{P}(A \cap B) = \frac{1}{6}$: This is the probability that both events occur. Notice that $A$ is a subset of $B$; therefore, $A \cap B = A$.

Intuitively, we expect $\mathbb{P}(A|B) = \frac{1}{3}$, and indeed we note that

$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

In general, it is true that

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

Why is this true?

(Informal Discussion.) Suppose that the sample space is $\Omega$ and we have two events $A, B \subset \Omega$. Then, consider the following Venn Diagram, where $\mathbb{P}(A)$ is represented by $\frac{\text{area}(A)}{\text{area}(\Omega)}$ and $\mathbb{P}(B)$ is represented by $\frac{\text{area}(B)}{\text{area}(\Omega)}$.
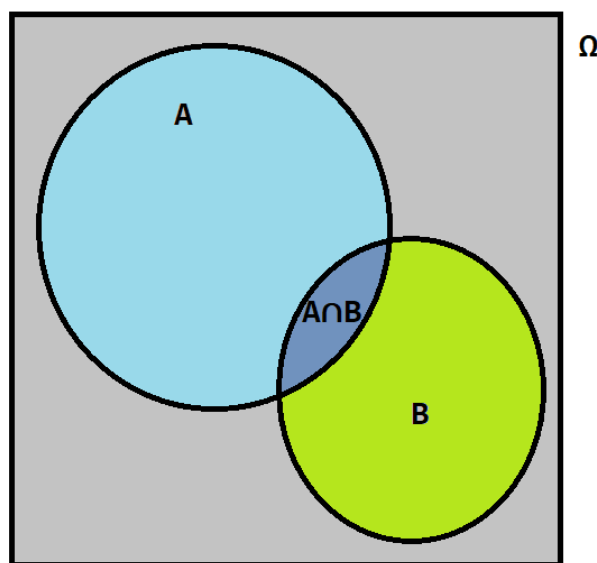


Now, the idea is that when we randomly throw a dart at the "dartboard" $\Omega$, it will "land in" $A$ with probability $\mathbb{P}(A)$ and similarly for $B$.

Now, suppose that we are told that the dart landed somewhere in $B$, but we don't know anything else beyond that. Then, since the dart was thrown randomly, it should be in some random position in $B$ (i.e. nowhere in $B$ should be more likely than anywhere else). Now, what is the probability that it landed in $A$, *given* that it landed somewhere in $B$?
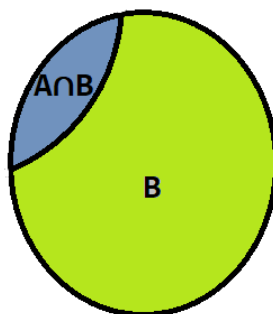
The answer is to find out what is the probability that the dart landed in the intersection, that is, in the $A \cap B$ region. In this case, we can consider their *ratios* – in this case, we consider the area of $A \cap B$ against the area of $B$. Then, in effect, $B$ becomes the new sample space (i.e. the "dartboard"), since we now know that the outcome of the experiment is in $B$. Therefore, we have

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

With this in mind, we can now give a formal definition.

> **Definition 4.2: Conditional Probability**
>
> Let $A$ and $B$ be two events. Suppose that $\mathbb{P}(B) > 0$. Then, the **conditional probability** of $A$, given that $B$ has occurred, is $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

Recall that $\mathbb{P}(\omega)$ is a probability distribution on the sample space $\Omega$ *if* we have

- $\mathbb{P}(\omega) \geq 0$ for all $\omega \in \Omega$, and

- $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$.

Note that the function $\mathbb{P}(\omega|B)$ is *also* a probability distribution, but now the *sample space* is $B$. In particular:

1. $\mathbb{P}(\omega|B) = \frac{\mathbb{P}(\omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\omega)}{\mathbb{P}(B)} \geq 0$ for all $\omega \in B$. Note that this is true since $\mathbb{P}(B) > 0$ and $\mathbb{P}(\omega) \geq 0$ for all $\omega \in \Omega$.

2. $\sum_{\omega \in B} \mathbb{P}(\omega|B) = \frac{1}{\mathbb{P}(B)} \sum_{\omega \in B} \mathbb{P}(\omega) = \frac{1}{\mathbb{P}(B)} \mathbb{P}(B) = 1$.

Note that, by multiplying both sides of the conditional probability formula $P(A|B)$ by $P(B)$, we get the following formula:

> **Theorem 4.1: Probability Chain Rule**
>
> $$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B).$$

**Remarks:**

- The conditional probability formula only holds when $\mathbb{P}(B) > 0$.

- The probability chain rule $\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B)$ holds even when $\mathbb{P}(B) = 0$.

(Example.) On sunny days, Vito goes for a walk with probability $\frac{4}{5}$. On rainy days, Vito goes for a walk with probability $\frac{1}{10}$. Suppose that, in San Diego, it rains on any given day with probability 3%. Find the probability that Vito goes for a walk today.

> We let $W$ be the event that Vito goes for a walk today; we want to find $\mathbb{P}(W)$. Let $S$ be the event that it is sunny today.
>
> We know that it rains with a 3% chance, so the change of it being sunny is 97%, or $\frac{97}{100}$. Thus, $\mathbb{P}(S) = \frac{97}{100}$. We also know that the probability that Vito goes for a walk, *given* that it is sunny[a], is $\mathbb{P}(W|S) = \frac{4}{5}$. Likewise, the probability that Vito goes for a walk, given that it is rainy (not a sunny day), is $\mathbb{P}(W|S^C) = \frac{1}{10}$.
>
> Recall that, by the Law of Total Probability, we have
>
> $$\mathbb{P}(W) = \mathbb{P}(W \cap S) + \mathbb{P}(W \cap S^C) \implies \mathbb{P}(W \cap S) = \mathbb{P}(W) - \mathbb{P}(W \cap S^C).$$
>
> Additionally, we know that
>
> $$\mathbb{P}(W \cap S^C) = \mathbb{P}(S^C)\mathbb{P}(W|S^C) = \frac{3}{100}\frac{1}{10} = \frac{3}{1000}.$$
>
> Therefore, applying this formula to the probability chain rule and solving, we get the following work:
>
> $$\mathbb{P}(W \cap S) = \mathbb{P}(S)\mathbb{P}(W|S) \qquad\qquad \text{Probability Chain Rule}$$
> $$\implies \mathbb{P}(W) - \mathbb{P}(W \cap S^C) = \mathbb{P}(S)\mathbb{P}(W|S) \quad \text{Applying Law of Total Probability}$$
> $$\implies \mathbb{P}(W) = \mathbb{P}(S)\mathbb{P}(W|S) + \mathbb{P}(W \cap S^C) \quad \text{Add } \mathbb{P}(W \cap S^C) \text{ to Both Sides}$$
> $$\implies \mathbb{P}(W) = \frac{97}{100}\frac{4}{5} + \frac{3}{1000} \qquad\qquad \text{Substitute Values}$$
> $$\implies \mathbb{P}(W) = \frac{779}{1000} \qquad\qquad\qquad \text{Simplify}$$
>
> ---
> [a]Note that we were not told that the probability of him going for a walk *and* it being sunny is $\frac{4}{5}$; all we're told is that *if* it is a sunny day, then he'll go for a walk with probability $\frac{4}{5}$.

## 4.3   Law of Total Probability, Conditional Version

Recall that, if $A \subset \Omega$ is an event, and if $B_1, \ldots, B_n$ are partitions of $\Omega$, then, the Law of Total Probability is given by

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i).$$

Using the probability chain rule, we have

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A|B_i).$$

This is the conditional version of the Law of Total Probability.

> **Theorem 4.2: Law of Total Probability, Conditional Version**
>
> Suppose that $A \subset \Omega$ is an event and that the events $B_1, \ldots, B_n$ partition that sample space $\Omega$. Then, we have that
>
> $$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A|B_i).$$

## 4.4   Independent Events

Now that we have defined conditional probability, we can formally define what it means for events to be independent.

> **Definition 4.3: Independent Events**
>
> Two events $A, B$ are **independent** if either
>
> 1. $\mathbb{P}(A) = 0$,
>
> 2. $\mathbb{P}(B) = 0$, or
>
> 3. $\mathbb{P}(A|B) = \mathbb{P}(A)$ and $\mathbb{P}(B|A) = \mathbb{P}(B)$, in the case that both $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$.

**Remark:** With regards to (3), the occurrence of $B$ does not affect the occurrence of $A$, and vice versa.

Now, by the probability chain rule, if two events $A$ and $B$ are independent, then

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

Indeed, $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B)$. This introduces the next theorem:

> **Theorem 4.3**
>
> If $A$ and $B$ are independent, then
> $$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Now, independence is more complicated when more than two events are involved. In particular:

> **Definition 4.4**
>
> Events $A_1, \ldots, A_n$ are **mutually independent** if, for any $A_{i_1}, \ldots, A_{i_k}$, we have that
>
> $$\mathbb{P}\left(\bigcap_{i=1}^{k} A_{i_k}\right) = \prod_{i=1}^{k} \mathbb{P}(A_{i_k}).$$

In particular, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbb{P}(A_i).$$

A weaker version of independence is as follows:

> **Definition 4.5**
>
> Events $A_1, \ldots, A_n$ are **pairwise independent** if each pair $A_i$, $A_j$ (with $i \neq j$) are independent.

Note that if $n > 2$, then mutually and pairwise independence are **not** equivalent.

So, in essense, the difference can be highlighted like so:

- Mutual independence: every event is independent of any intersection of the other events.

- Pairwise independence: any two events are independent.

## 4.5   Joint Probability Distributions

We now talk about joint probability distributions.

### 4.5.1 Joint PMFs

Suppose that $X_1, \ldots, X_n$ are discrete random variables with PMFs

$$p_{X_i} = \mathbb{P}(X_i = x).$$

Then, the joint PMF of the RV $\mathbf{X} = (X_1, \ldots, X_n)$ is the function

$$p_{\mathbf{X}}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

for all $x_1, \ldots, x_n$. Here, we can think of this as a function from a sample space to $\mathbb{R}^n$ (higher dimensions).

Often, it can be quite difficult to find the joint PMFs. However, in the special case that $X_1, \ldots, X_n$ are **independent**, it is just the product of the individual PMFs; that is,

$$p_{\mathbf{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

We say that $X_1, \ldots, X_n$ are **independent and identically distributed**[3] (IID for short) if they are mutually independent, and all have the same distribution.

> (Example.) Consider a sequence $X_1, \ldots, X_n$ of $n$ Bernoulli($p$) trials. In this case, each trial has the simple distribution
> $$\mathbb{P}(X_i = 1) = p$$
> and
> $$\mathbb{P}(X_i = 0) = 1 - p.$$

### 4.5.2 Joint PDFs

In general, though, in such a process, $X_i$ can have any given distribution. In any case, if $X_1, \ldots, X_n$ are IIDs, then their joint PDF takes the simple form known as a **product distribution**. In particular, since the $X_i$'s are IIDs, they have the same PDFs $p(x) = \mathbb{P}(X_i) = x$ no matter what $i$ is. Hence, the joint PDF is just

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i).$$

> (Example.) For example, in a sequence of Bernoulli($p$) trials,
> $$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = p^k (1-p)^{n-k},$$
> if exactly $k$ of the $x_i = 1$ and the other $n - k$ of the $x_i = 0$.

> (Example Problem.) Let $X, Y$ be independent Bernoulli($1/2$) random variables. Let $Z$ be a random variable that takes the value 1 if $X + Y = 1$, and 0 otherwise. Show that $X, Y, Z$ are pairwise, but not mutually, independent.

---

[3]This is known as **independent trial process** in the textbook.

Trivially, we know that $\mathbb{P}(X = 1) = \mathbb{P}(X = 0) = \frac{1}{2}$ and $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = \frac{1}{2}$, since these are just Bernoulli$(1/2)$ random variables. We know that $X$ and $Y$ are both independent, so the values that they take do not affect each other. Using a table, we can fill out some possible values of $X$, $Y$, and $Z$.

| **X** | **Y** | **X + Y** | **Z** |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 2 | 0 |

Since $X$ and $Y$ are independent, it follows that

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}.$$

Recall that random variables $X_1, \ldots, X_n$ are independent if

$$p_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n).$$

From the table above, we know that

$$\mathbb{P}(Z = 1) = \mathbb{P}(Z = 0) = \frac{1}{2}.$$

Now, recall that

$$p_{X,Z}(x, z) = \mathbb{P}(X = x, Z = z).$$

For any selection of $(x, z) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, we have

$$p_{X,Z}(x, z) = \frac{1}{4}.$$

Likewise, we have that

$$p_X(x) = \frac{1}{2}$$

and

$$p_Z(z) = \frac{1}{2}.$$

Therefore,

$$p_{X,Z}(x, z) = \frac{1}{4} = \left(\frac{1}{2}\right)^2.$$

However, note that we simply have $p_{X,Z}(x, z) = p_X(x)p_Z(z)$, so it follows that $X$ and $Z$ are pairwise independent. The same argument can be made for $Y$ and $Z$. Therefore, $X, Y, Z$ are pairwise independent.

Now, note that

$$p_{X,Y,Z}(0, 1, 1) = \frac{1}{4}.$$

But,

$$p_X(0)p_Y(1)p_Z(1) = \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}.$$

Since $p_{X,Y,Z}(0, 1, 1) \neq p_X(0)p_Y(1)p_Z(1)$, clearly $X, Y, Z$ are not mutually independent.

## 4.6   Bayes' Formula

Bayes' Formula is a powerful – and very famous – application of the conditional probability formula. Often, it can be difficult to calculate a conditional probability $\mathbb{P}(A|B)$ of interest. However, the other way around, $\mathbb{P}(B|A)$, could be easier to find. Bayes' Formula gives us a way of finding $\mathbb{P}(A|B)$, provided that we know $\mathbb{P}(A)$, $\mathbb{P}(B)$, and $\mathbb{P}(B|A)$. Recall that the conditional probability formula is given by

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

We also know that

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

So, solving for $\mathbb{P}(A \cap B)$, we have

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

Then, setting these terms equal to each other, we have

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

Thus, we get

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

It is often useful to apply the conditional Law of Total Probability in the denominator; that is,

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

So, sometimes, Bayes' Rule is stated like so:

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^C)\mathbb{P}(B|A^C)}}.$$

Now, there is a general version of Bayes' Rule; suppose that $B \subset \Omega$ is an event and $A_1, \ldots, A_n$ partitions the sample space $\Omega$. Then, for each $1 \leq j \leq n$, we have

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

Here[4],

- The $\mathbb{P}(A_j)$ are called **prior probabilities**.

- The $\mathbb{P}(A_j|B)$ are called **posterior probabilities**.

- The events $A_j$ are called **hypotheses**.

- The event $B$ is called the **evidence**.

Often, we want to see which hypothesis is more likely given the evidence.

(Example.) Suppose that a doctor gives a patient a test for cancer. Before the test, all we know is that, on average, 1 in every 1000 women develop this cancer. According to the manufacturer, this test is 99% accurate at detecting the cancer, when it is there. However, there is a 5% chance it will show positive when the cancer is not there (i.e. there is a 1% chance of a false negative, and a 5% chance of a false positive). To make things more clear:

- If she has the cancer, there's a 1% chance the test will incorrectly say negative, and a 99% chance the test will correctly say positive.

---

[4]In this course, we aren't expected to memorize these.

- If she does not have the cancer, there's a 5% chance the test will incorrectly say positive.

Now, suppose that the patient tests positive. *With what probability does she actually have the cancer?*

> Let $C$ be the event that she has this cancer. Let $P$ be the event that this test is positive. We want to find $\mathbb{P}(C|P)$. We know that
>
> $$\mathbb{P}(C) = 0.001 \qquad \mathbb{P}(P|C) = 0.99 \qquad \mathbb{P}(P|C^C) = 0.05.$$
>
> By Bayes' Rule, we have
>
> $$\mathbb{P}(C|P) = \frac{\mathbb{P}(C)\mathbb{P}(P|C)}{\mathbb{P}(C)\mathbb{P}(P|C) + \mathbb{P}(C^C)\mathbb{P}(P|C^C)} = \frac{0.001(0.99)}{0.001(0.99) + 0.999(0.05)} \approx 1.94\%.$$

**Remark:** Given that the test is supposed to be 99% accurate and that the patient tested positive, one would think that the patient has the cancer; so, a near 2% probability that the patient has cancer is quite surprising. However, there are a few reasons why this may be the case.

1. Even though the test is fairly accurate when you do have the cancer, when you don't have the cancer there is a decent chance (5%) that it will be incorrect.

2. This cancer is *extremely* rare; it's so rare that there is a much better chance that this test will incorrectly read positive than if you actually have cancer.

Now, without any additional information (i.e. without the test), we could only assume that the patient had this cancer with probability 0.1%. After testing positive, this probability increases by quite a bit, but it is still only 1.94%. This example demonstrates that it is very difficult to design an accurate test for a rare disease; in this example, the probability of a false positive is much more likely than the patient actually having cancer.

> (Example Problem.) An airplane is missing. Based on its flight plan, it has been determined that the airplane is equally likely to be in any one of three locations (each with probability 1/3). Some locations are easier to search than others for various geological reasons. If the airplane is in location $1, 2, 3$ then the team will find it while searching there with probability 1/2, 1/3, 1/4, respectively. Suppose that location 1 is searched first, and the airplane is not found there. Show that, given this, with conditional probability 2/5 the airplane is in location 2.

We will use Bayes' Formula,

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

where $A$ is the event that the plane is in location 2 and $B$ is the event that the airplane was not found in location 1. We note that

$$\mathbb{P}(\text{Plane in Loc. 1}) = \boxed{\mathbb{P}(\text{Plane in Loc. 2}) = \mathbb{P}(A)} = \mathbb{P}(\text{Plane in Loc. 3}) = \frac{1}{3}.$$

Let $L_i$ be the event that the plane is in location $i$. By the conditional law of total probability, we know that

$$\mathbb{P}(B) = \mathbb{P}(L_1)\mathbb{P}(B|L_1) + \mathbb{P}(L_2)\mathbb{P}(B|L_2) + \mathbb{P}(L_3)\mathbb{P}(B|L_3).$$

We now consider each case individually:

- Suppose the plane was in location 1, but the search occurred (and failed) in location 1. Since there was a $\frac{1}{2}$ chance that the plane could be found in location 1, then there is a $1 - \frac{1}{2} = \frac{1}{2}$ chance that the plane will not be found in location 1. So:

$$\mathbb{P}(L_1)\mathbb{P}(B|L_1) = \frac{1}{3}\left(1 - \frac{1}{2}\right) = \frac{1}{6}.$$

- Suppose the plane was in location 2, but the search occurred in location 1. Note that if the plane was in location 2, then trivially there is a 100% chance the plane will not be found in location 1, so it follows that $\mathbb{P}(B|L_2) = \mathbb{P}(B|A) = 1$. Then:

$$\mathbb{P}(L_2)\mathbb{P}(B|L_2) = \frac{1}{3} \cdot 1 = \frac{1}{3}.$$

- Suppose the plane was in location 3, but the search occurred in location 1. The same reason from the previous part can be applied to this part as well. Then:

$$\mathbb{P}(L_3)\mathbb{P}(B|L_3) = \frac{1}{3} \cdot 1 = \frac{1}{3}.$$

Thus, $\mathbb{P}(B) = \frac{5}{6}$. So, we have

$$\begin{aligned}
\mathbb{P}(A|B) &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \\
&= \frac{1 \cdot 1/3}{5/6} \\
&= \frac{1}{3} \cdot \frac{6}{5} \\
&= \boxed{\frac{2}{5}}.
\end{aligned}$$

Thus, we got the desired answer.

## 4.7 Conditional Probability: Continuous Case

Conditional probability is slightly different in the "world" of continuous random variables. Most of the same definitions carry over from the discrete case, but since we're dealing with *PDFs*[5] $f(x)$ instead of PMFs $p(x) = \mathbb{P}(X = x)$ (which are probabilities, unlike PDFs).

---

[5]Remember, PDFs are not probabilities; they are densities. They allow us to get a probability by integrating, but they're not probabilities.

Recall that if $\mathbb{P}(\omega)$ is a discrete probability distribution and $\mathbb{P}(B) > 0$, then the distribution $\mathbb{P}(\omega|B) = \frac{\mathbb{P}(\omega)}{\mathbb{P}(B)}$ is a probability distribution on $B$.

---

**Definition 4.6**

Let $X$ be a continuous random variable with PDF $f$. Suppose that $B$ is an event with $\mathbb{P}(B) > 0$. Then, the conditional PDF of $X$, given $B$, is

$$f(x|B) = \begin{cases} \frac{f(x)}{\mathbb{P}(B)} & x \in B \\ 0 & x \notin B \end{cases}.$$

---

We note that $f(x) \geq 0$ and

$$\int_{-\infty}^{\infty} f(x|B)dx = \frac{1}{\mathbb{P}(B)} \int_B f(x)dx = \frac{1}{\mathbb{P}(B)}\mathbb{P}(B) = 1.$$

Note that the integration over a region gives us the probability of being in that region, i.e. integrating $f(x)$ over the region $B$ gives us just $\mathbb{P}(B)$. Therefore, $f(x|B)$ is indeed a probability density on $B$.

---

(Example.) Recall the spinner example. The location of the spinner, when it eventually comes to rest, is uniform on the unit circle; thus,
$$f(x) = 1 \text{ for } x \in [0, 1).$$

Suppose that the spinner comes to rest in the upper-half of the circle, i.e. in $[0, 1/2]$. What is the (conditional) probability with which it lands in the region $[1/6, 2/3]$?

> The probability of $B$ occurring (the event that the spinner lands in the upper-half of the circle) is $\frac{1}{2}$ since it is uniform from $[0, 1)$.
>
> Normally, we would integrate from $\frac{1}{6}$ to $\frac{2}{3}$. However, since we know that the spinner lands in the region $[0, 1/2]$, we know that there is no density *outside* of this region. Since $x$ is outside of the region for $\frac{1}{2} \leq x \leq \frac{2}{3}$, we don't need to consider it. Instead, we only need to consider the region $[1/6, 1/2]$. Hence,
> $$\int_{\frac{1}{6}}^{\frac{1}{2}} \frac{1}{1/2}dx = \frac{2/6}{1/2} = \frac{2}{3}.$$

---

Recall that the CDF of a continuous random variable is the function

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)du.$$

---

**Definition 4.7**

A sequence of random variables $X_1, \ldots, X_n$ is said to be **mutually independent** if their joint CDF is the product of the individual CDF. That is, if

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} F_i(x_i)$$

for *all* $x_1, \ldots, x_n$.

---

By calculus, we can find the joint PDF of a sequence of continuous random variables by differentiating:

$$f(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1, \ldots, \partial x_n} F(x_1, \ldots, x_n).$$

From this, we have the following theorem:

> **Theorem 4.4**
>
> $X_1, \ldots, X_n$ are mutually independent if and only if their joint PDF is the product of the individual PDFs, $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$.

Additionally, consider the following theorem:

> **Theorem 4.5**
>
> If $X_1, \ldots, X_n$ are mutually independent, and $f$ is a continuous function, then $f(X_1), \ldots, f(X_n)$ are mutually independent.

---

(Example.) Suppose that $X_1$ and $X_2$ are independent exponential random variables with rates $\lambda_1$ and $\lambda_2$. Find the PDF of $M = \min\{X_1, X_2\}$.

> Recall that if $X$ is an exponential rate $\lambda$ random variable if and only if its survival function is
>
> $$S(x) = 1 - F(x) = P(X > x) = e^{-\lambda x}.$$
>
> Since $X_1, X_2$ are independent, their joint PDF is the product (from differentating)
>
> $$f(x_1, x_2) = f_1(x_1) f_2(x_2) = \lambda_1 e^{-\lambda_1 x_1} \cdot \lambda_2 e^{-\lambda_2 x_2}.$$
>
> Also, note that $M = \min\{X_1, X_2\} > m$ if and only if $X_1 > m$ and $X_2 > m$. Hence,
>
> $$\begin{aligned}
> \mathbb{P}(M > m) &= \int_m^\infty \int_m^\infty f(x_1, x_2) dx_1 dx_2 \\
> &= \left( \int_m^\infty \lambda_1 e^{-\lambda_1 x_1} dx_1 \right) \left( \int_m^\infty \lambda_2 e^{-\lambda_2 x_2} dx_2 \right) \\
> &= \mathbb{P}(X_1 > m) \mathbb{P}(X_2 > m) \\
> &= e^{-\lambda_1 m} e^{-\lambda_2 m} \\
> &= e^{-(\lambda_1 + \lambda_2) m}.
> \end{aligned}$$
>
> Therefore, $M = \min\{X_1, X_2\}$ is an exponential random variable with rate $\lambda_1 + \lambda_2$ (the sum of the rates of $X_1$ and $X_2$).

## 4.8   Memoryless Property of Exponential Variables

> **Theorem 4.6**
>
> Suppose $X$ is exponential with rate $\lambda$. Then, $X$ has no memory, in the sense that for any $t, s > 0$,
>
> $$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t).$$

That is, given the *survival* of $X$ to time $s$, it then behaves like a "brand new" exponential rate $\lambda$ random variable.

> *Proof.* Note that $\mathbb{P}(X > x) = e^{-\lambda x}$ for any $x > 0$. Hence,
>
> $$\mathbb{P}(X > s + t | X > s) = \int_{s+t}^\infty \frac{\lambda e^{-\lambda x}}{e^{-\lambda s}} dx = e^{\lambda s} \mathbb{P}(X > s + t) = e^{\lambda s} e^{-\lambda(s+t)} = e^{-\lambda t}.$$

Hence, this is equal to $\mathbb{P}(X > t)$ as claimed. □

---

(Example: Beta Distribution.) The Beta$(\alpha, \beta)$ random variable, with parameters $\alpha, \beta > 0$, has PDF

$$f(x) = B(\alpha, \beta, x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

for $x \in [0,1]$ (and $f(x) = 0$ otherwise), where

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

is the beta integral. Now, when $\alpha, \beta$ are integers, then

$$B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}.$$

Note that when $\alpha = \beta = 1$, then $B(\alpha, \beta, x) = 1$. Thus, Beta$(1,1)$ is uniform on $[0,1]$.

---

The beta distribution is used in **Bayesian statistics**.

---

(Example: Drug Testing.) Suppose that a drug has an unknown probability $X$ of being effective. Therefore, the first time it is administered, it would be quite natural to assume that the distribution of $X$ is uniform on $[0,1]$. However, as time goes on, and more data is collected, we might want to update this distribution. For example, if it is successful in all except 6 of the first 100 trials, a uniform distribution would no longer seem appropriate.

Suppose that we given this drug to $n$ patients. Assuming independence, we can model this as a series of $n$ Bernoulli trials with (unknown) success probability $X$. If $X = x$, then the probability that the $i$ trials will be successful is the binomial probability

$$b(n, x, i) = \binom{n}{i}x^i(1-x)^{n-i}.$$

Hence, the conditional PMF is $p(i|x) = b(n, x, i)$.

If $X$ is uniform on $[0,1]$, which has PDF of 1 on $[0,1]$, then the $i$ of the $n$ trials will be successful with probability

$$\int_0^1 1 \cdot p(i|x)dx = \int_0^1 b(n, x, i)dx = \binom{n}{i}\int_0^1 x^i(1-x)^{n-i}dx.$$

We note that

$$\int_0^1 x^i(1-x)^{n-i}dx = B(i+1, n-i+1).$$

Hence, the PMF is given by

$$p(i) = \binom{n}{i}B(i+1, n-i+1).$$

Now, still assuming that $X$ is uniform, note that the joint distribution is given by $f(x, i) = p(i|x) \cdot 1 = b(n, x, i)$. So, the conditional PDF is

$$f(x|i) = \frac{f(x,i)}{p(i)} = \frac{b(n, x, i)}{\binom{n}{i}B(i+1, n-i+1)} = \frac{x^i(1-x)^{n-i}}{B(i+1, n-i+1)},$$

which is the PDF of a Beta$(i+1, n-i+1)$ random variable.

In fact, we could continue to update this distribution as we go along and continue to collect more data. So, for the first patient, we have seen 0 successes and 0 failures (no data yet). So, we assume that the probability of success on the first patient is distributed as a $\text{Beta}(1, 1) = \text{Uniform}[0, 1]$ random variable.

# 5    Distributions and Densities

We are now interested in looking at important probability distributions, along with their applications.

## 5.1    Types of Discrete Probability Distributions

The following are some of the most important discrete probability distributions, some of which we've seen before.

- Uniform

- Bernoulli/Indicator

- Binomial

- Geometric

- Negative Binomial

- Poisson

- Hypergeometric

### 5.1.1    Uniform Distribution

The idea is that we put equal mass on every element in the set. More formally:

> **Definition 5.1: Uniform Distribution**
>
> $X$ is **uniform** on a finite set $A$ if $p(x) = \frac{1}{|A|}$ for all $x \in A$.

**Remark:** This distribution has the special property that, for all $B \subset A$, $P(X \in B) = \frac{|B|}{|A|}$.

### 5.1.2    Bernoulli Distribution

Essentially, we can only ever get two values: a 1 or a 0, with probability $p$ and $1 - p$. More formally:

> **Definition 5.2: Bernoulli Distribution**
>
> $X$ is **Bernoulli**$(p)$ if $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

**Remark:** In terms of Bernoulli, we usually think of it as a trial (where there's either a success or failure).

> (Example: Indicator Random Variable.) Let $E$ be an event. The random variable $\mathbf{1}_E$ is equal to 1 if $E$ occurs and equal to 0 if $E^C$ occurs. This is known as a **indicator** of $E$. Note that $\mathbf{1}_E$ is a Bernoulli random variable with parameter $p = \mathbb{P}(E)$.

### 5.1.3    Binomial Distribution

The binomial distribution involves Bernoulli trials.

### Definition 5.3: Binomial Distribution

Suppose that $n$ independent Bernoulli($p$) trials are performed. Then, the number of successes observed during these $n$ trials has the **Binomial**$(n, p)$ distribution with

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

for $0 \leq k \leq n$.

**Remark:** In particular, for $k$ successful trials, the idea is that we pick $k$ trials to be successful (out of the $n$ trials). Then, we have probability $p^k$ for those trials to be successful, and probability $(1-p)^{n-k}$ for those trials to be failures. By independence, we can just multiply them out.

#### 5.1.4 Geometric Distribution

The geometric distribution also involves Bernoulli trials.

### Definition 5.4: Binomial Distribution

Suppose that we keep performing independent Bernoulli($p$) trials until the first success is observed. Then, the total number of trials performed has the **Geometric**$(p)$ distribution, with

$$p(k) = p(1-p)^{k-1}$$

for $k \geq 1$.

**Remarks:**

- Here, the $k$th trial is the successful trial, and all of the $k - 1$ trials are thus failures.

- A related distribution is known as the **Shifted Geometric**$(p)$ distribution. Instead of asking how many trials before the first success, we're asking how many *failures* before the first success. This has PMF
$$p(k) = p(1-p)^k$$
for $k \geq 0$.

- Here, $X$ is geometric if and only if $X - 1$ is shifted geometric.

The geometric random variable is, in a sense, the discrete analogue of the (continuous) exponential random variable. Recall that the exponential random variable is the only continuous random variable with the *memoryless property*. The geometric random variable *also* has this property, and is the only *discrete random variable* that does.

*Proof.* Suppose that $X$ is Geometric($p$). Then[a],

$$\mathbb{P}(X > k) = \sum_{\ell=k+1}^{\infty} p(1-p)^{\ell-1}$$
$$= p(1-p)^k \sum_{\ell=0}^{\infty} (1-p)^\ell$$
$$= p(1-p)^k \frac{1}{1-(1-p)}$$
$$= p(1-p)^k \frac{1}{p}$$
$$= (1-p)^k.$$

Hence,

$$\mathbb{P}(X > k + \ell | X > k) = \frac{\mathbb{P}(X > k + \ell)}{\mathbb{P}(X > k)}$$
$$= \frac{(1-p)^{k+\ell}}{(1-p)^k}$$
$$= (1-p)^\ell$$
$$= \mathbb{P}(X > \ell).$$

Thus, $X$ has the memoryless property.                                           □

---

[a] $X$ is the number of trials until the first success. Intuitively, we note that $X > k$ occurs if and only all of the first $k$ trials were failures.

### 5.1.5   Negative Binomial Distribution

The negative binomial distribution is a generalization of the geometric distribution. Instead of waiting for the first success, we wait for the $k$th success.

> **Definition 5.5**
>
> Suppose that we keep performing independent Bernoulli($p$) trials until the $k$th success is observed. Then, the total number of trials performed has the **Negative Binomial**$(k, p)$ distribution, with
>
> $$p(n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$
>
> for $n \geq k$.

**Remark:** Why is this the PMF?

- For the $p^k(1-p)^{n-k}$ part, there are $k$ successes, and there are $n - k$ failures like usual.

- For the binomial coefficient, the idea is that if the $n$th trial is the $k$th success, then there's no choice as to when the $k$th success will be; it has to be $\binom{n}{n}$. But, what about the $n - 1$ trials, of which $k - 1$ of them are successes? Well, there were exactly $k - 1$ successes during the first $n - 1$ trials. These could have occurred in any of $\binom{n-1}{k-1}$ ways.

### 5.1.6   Poisson Distribution

This is the most important distributions in all of probability theory.

> **Definition 5.6**
>
> A **Poisson** RV with rate $\lambda > 0$ has PMF
>
> $$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$
>
> for $k \geq 0$.

**Remark:** One way to remember this PMF is to recall the Taylor expansion $e^\lambda = \sum_{k=0}^\infty \frac{\lambda^k}{k!}$, which can be used to show that this is indeed a PMF, $\sum_{k=0}^\infty p(k) = 1$.

The Poisson is related to the exponential and binomial distributions. In particular, the connection with the binomial is that if $N$ is the Binomial$(n, \lambda/n)$, then as $n \mapsto \infty$, it "converges" to Poisson$(\lambda)$. Note, if

$$\mathbb{P}(N = k) = \binom{n}{k} (\lambda/n)^k (1 - \lambda/n)^{n-k}.$$

For any fixed integer $k \geq 0$, we have

$$\binom{n}{k} \approx n^k/k!$$

and

$$(1 - \lambda/n)^{n-k} \approx e^{-\lambda},$$

so it follows that

$$\mathbb{P}(N - k) \approx e^{-k}\frac{\lambda^k}{k!},$$

which is the PMF of a Poisson($\lambda$). Thus, since a Binomial($n, p$) is approximately Poisson($\lambda$), when $p \approx \lambda/n$, the Poisson is useful for studying the occurrence of **rare events**.

If $p$ is small compared with $n$, then

$$\boxed{\mathbb{P}(\text{Binomial}(n, p) = k) \approx \mathbb{P}(\text{Poisson}(np) = k)}.$$

A good "rule of thumb" is $n \geq 100$ and $np \leq 10$ in order for the approximation to be reasonable.

---

(Example.) On average, a typist makes one typo in every 1000 words. Approximate the probability of having at most 2 typos on the first 3 pages ($\approx 300$ words) of a manuscript they have typed.

> We could use a Binomial($300, 0.001$) random variable here, but it would lead to the somewhat length calculation
>
> $$\sum_{k=0}^{2} \binom{300}{k}(0.001)^k(0.999)^{300-k} \approx 99.6429\%.$$
>
> Here, the calculation is essentially:
>
> - The probability that there is at most 0 typos, and
>
> - The probability that there is at most 1 typo, and
>
> - The probability that there is at most 2 typos.
>
> Note that $n = 300$ and $p = \frac{1}{1000}$, so $n = 300 > 100$ and $np = 300(0.001) = 0.3 < 10$, so it follows that we will get a good approximation; in particular, we expect a Poisson($0.3$) random variable $X$ to give a good approximation. Indeed,
>
> $$\mathbb{P}(X \leq 2) = \sum_{k=0}^{2} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{2}\frac{\lambda^k}{k!} = e^{-0.3}\left(1 + 0.3 + \frac{(0.3)^2}{2}\right) \approx 99.6401\%.$$
>
> This is nearly as good as what we found by using the Binomial.

---

### 5.1.7   Hypergeometric Distribution

This distribution is useful for sampling *without* replacement.

---

(Example.) Suppose that a lake contains $N$ fish, where $k$ of them are big and $N - k$ of them are small. Suppose that a fisherman catches $n$ fish. Let $X$ be the number of them that are big. Then, with probability

$$\mathbb{P}(X = x) = \frac{\binom{k}{n}\binom{N-k}{n-x}}{\binom{N}{n}},$$

where $x$ of them will be big.

---

Note that if we instead sampled *with* replacement, then the probability that $k$ of the $n$ fish are big would just be the Binomial probability

$$b(n, k/N, x) = \binom{n}{x}(k/N)^x(1 - k/N)^{n-x}.$$

If both of $N$ and $k$ are much larger than $n$, then there is not much difference. Intuitively, because then there is little chance that we would catch the same fish twice anyhow, so sampling with or without replacement would be essentially the same.

### 5.1.8 Benford Distribution

In many "real life" datasets, the first digits tend to **not** be uniform on $\{1, 2, \ldots, 9\}$. In particular, digit 1 tends to be the most likely to occur. The **Benford distribution**, in many cases, does a better job of modeling this distribution of the leading digits occurring in "typical" data.

---

(Example.) The Benford distribution for the *first* digit (base 10) has PMF

$$p(k) = \log_{10}(1 + 1/k)$$

for $1 \leq k \leq 9$.

---

In particular, the first digit of most "typical" datasets is a 1 with about a 30% chance.

## 5.2 Types of Continuous Probability Distributions

The following are some of the most important continuous probability distributions, some of which we've seen before.

- Uniform

- Exponential

- Gamma

- Normal

- Cauchy

### 5.2.1 Uniform Distribution

> **Definition 5.7: Uniform Distribution**
>
> $U$ has **Uniform**$[a, b]$ distribution, for $a < b$, if its PDF is
>
> $$f(u) = \frac{1}{b - a}$$
>
> for $a \leq u \leq b$.

**Remark:** Note that $b - a$ is the *length* of $[a, b]$.
**Note:** There are also higher-dimensional uniform distributions, but then we replace length with area of volume.

### 5.2.2 Exponential & Gamma Distribution

> **Definition 5.8: Exponential Distribution**
>
> $X$ is Exponential($\lambda$) with rate $\lambda > 0$ if its PDF is
>
> $$f(x) = \lambda e^{-\lambda x}$$
>
> for $x > 0$.

Note that there is an important connection between the Exponential and the Poisson, which we will now describe.

---

(Example: Busy Server.) Suppose that a single server queue (e.g. call center, bank, etc.) is very busy, so that there is always someone in the queue. Suppose that service times are independent and Exponential($\lambda$)[a]. As soon as someone has been served, the next person in the queue starts being served immediately. Let $X_1, X_2, \ldots$ be an IID sequence of Exponential($\lambda$) random variables. Then, the time $T_n$ at which the point the $n$th person has been served is distributed as

$$\sum_{i=1}^{n} X_i.$$

This sum of $n$ IID Exponential($\lambda$) random variables has a special distribution, called the **Gamma**$(n, \lambda)$ distribution. This has PDF[b]

$$g(x) = \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!}$$

for $x > 0$.

---
[a]As you wait longer, the time that you wait "decays," or decreases.
[b]When $n = 1$, this reduces to $\lambda e^{-\lambda x}$, the PDF of a single Exponential($\lambda$).

---

The Poisson distribution arises when we ask: What is the distribution for the number $N_t$ of people served by time $t > 0$? At any given time $t > 0$, we will (with probability 1, since the service times are continuous) be in the middle of serving someone. This person does not count towards $N_t$.

Note that $(N_t : t > 0)$ is called the **Poisson process**[6]. This is a *collection*, indexed by time, of random variables[7] In particular, the random variable $N_t$ has the Poisson($\lambda t$) distribution.

---

*Proof.* Note that $N_t = k$ if and only if the $k$th person is served at some time $T_k = s \leq t$, and then the next service $X_{k+1} > t - s$. In other words, we need to have finished serving $k$ people and be in the middle of serving the $(k+1)$th person. Since

$$T_k = \sum_{i=1}^{k} X_i$$

and $X_{k+1}$ are independent, it follows that

$$\mathbb{P}(N_t = k) = \int_0^t f_{T_k}(s)[1 - F_{X_{k+1}}(t-s)]ds.$$

---
[6]This is a fascinating mathematical object with many properties and applications, which won't be covered here.
[7]Such an object is called a **stochastic process**.

Since $T_k$ is Gamma$(k, \lambda)$ and $X_{k+1}$ is Exponential$(\lambda)$, it follows that

$$
\begin{aligned}
\mathbb{P}(N_t = k) &= \int_0^t f_{T_k}(s)[1 - F_{X_{k+1}}(t - s)]ds \\
&= \int_0^t \lambda e^{-\lambda s} \frac{(\lambda s)^{k-1}}{(k-1)!} \cdot e^{-\lambda(t-s)} ds \\
&= e^{-\lambda t} \frac{\lambda^k}{(k-1)!} \int_0^t s^{k-1} ds \\
&= e^{-\lambda t} \frac{(\lambda t)^k}{k!}.
\end{aligned}
$$

Hence, this is the PMF of a Poisson, as claimed. □

### 5.2.3 Normal/Gaussian Distribution

Recall that a Binomial converges to a Poisson if

$$
p = \lambda/n \mapsto 0
$$

as

$$
n \mapsto \infty.
$$

On the other hand, if $p$ is *fixed* (not converging to 0), the Binomial approaches a different distribution as $n \mapsto \infty$ called the **Normal** or **Gaussian** distribution. Indeed, as $n$ goes to infinity, we see a bell-shaped curve.

---

**Definition 5.9: Normal Distribution**

$X$ is **Normal**$(\mu, \sigma^2)$ if its PDF is

$$
\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
$$

for $-\infty < x < \infty$.

---

Here, $\mu$ is the "center" of the density and $\sigma$ is the measure of the "spread" of the density.
When $\mu = 0$ and $\sigma^2 = 1$, $X$ is called a standard normal, and its PDF is given the special notation

$$
\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.
$$

If $X$ is not standard, then you can "standardize" by taking $Z = (X - \mu)/\sigma$. Then, $X$ is Normal$(\mu, \sigma^2)$ if and only if $Z$ is Normal$(0, 1)$.

### 5.2.4 Cauchy Distribution

Now, suppose that $X$ and $Y$ are two independent standard Normal random variables. A very interesting distribution arises if we consider the ratio

$$
Z = X/Y.
$$

Since $X$ and $Y$ are independent,

$$
f_Z(z) = \int_{S_z} f_{X,Y}(x, y) dx dy,
$$

where $S_z = \{\{(x, y) \mid x/y = z\}$. We make a change of variables $x = uz$ and $y = u$. Then, as $u$ varies over $\mathbb{R}$, we get the whole set $S_z$. The Jacobian of this transformation is $|u|$, so

$$
f_Z(z) = \int_{S_z} f_{X,Y}(x, y) dx dy = \int_{-\infty}^{\infty} |u| f_{X,Y}(uz, u) du.
$$

This is the same as

$$2\left(\frac{1}{\sqrt{2\pi}}\right)^2 \int_0^\infty u e^{-(uz)^2/2 - u^2/2} du.$$

It can be shown that

$$\int_0^\infty x e^{-cx^2} dx = \frac{1}{2c}.$$

Hence,

$$f_Z(z) = \frac{1}{\pi} \int_0^\infty u e^{-\frac{u^2(1+z^2)}{2}} du = \frac{1}{\pi(1+z^2)}.$$

A random variable with this PDF is called a (standard) **Cauchy** random variable.

Note that the Cauchy distribution has some interesting properties. In partivular, it has no expected/average value. So, if you take an IID sequence $X_1, X_2, \ldots$ of Cauchy random variables, the limit

$$\lim_{n \mapsto \infty} \frac{1}{n} \sum_{i=1}^n X_i$$

does not exist.

## 5.3 CDF/PDF Transformations

We can use the distributions that we've talked about to build more *complex* distributions.

---

### Theorem 5.1: CDF Transformation Theorem

Let $X$ be a continuous random variable and $\Phi$ is a strictly monotone function. Then, the random variable

$$Y = \Phi(X)$$

has CDF

1. $F_Y(y) = F_X(\Phi^{-1}(y))$, if $\Phi$ is increasing.

2. $F_Y(y) = 1 - F_X(\Phi^{-1}(y))$, if $\Phi$ is decreasing.

---

*Proof.* Suppose that $\Phi$ is strictly increasing. Then, the inverse function $\Phi^{-1}$ exists and is increasing. Therefore,

$$\Phi(X) \leq y$$

if and only if

$$\Phi^{-1}(\Phi(X)) = X \leq \Phi^{-1}(y).$$

Hence,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\Phi(X) \leq y) = \mathbb{P}(X \leq \Phi^{-1}(y)) = F_X(\Phi^{-1}(y)),$$

as claimed. The strictly decreasing case is similar. $\square$

Note that, by differentiating using the Calculus Chain Rule, we obtain the following corollary.

---

### Corollary 5.1: PDF Transformation Theorem

Let $X$ be a continuous random variable and $\Phi$ a strictly monotone function. Then, the random variable $Y = \Phi(X)$ has PDF

$$f_Y(y) = f_X(\Phi^{-1}(y)) \left| \frac{d}{dy} \Phi^{-1}(y) \right|.$$

---

Note that there is a transformation theorem in the case of discrete random variables, but it is much easier. In particular, the PMF of a random variable $Y = \Phi(X)$ is simply the function

$$p_Y(y) = \sum_{x:\Phi(x)=y} p_X(x).$$

---

(Example.) Let $U$ be Uniform on $[0, 1]$. Then, consider the transformed version

$$V = 1 - U,$$

which is also uniform on $[0, 1]$. Note that

$$\Phi(u) = 1 - u$$

is decreasing, and $\Phi^{-1}(v) = 1 - v$. Therefore, by the CDF Transformation Theorem, we have

$$F_V(v) = 1 - F_U(\Phi^{-1}(v)) = 1 - F_U(1 - v) = 1 - (1 - v) = v.$$

Hence, $V$ is Uniform on $[0, 1]$.

---

(Example.) Let $X$ be a standard Normal$(0, 1)$. Consider

$$Y = X^2.$$

Recall that $X$ has PDF

$$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

The function

$$\Phi(x) = x^2$$

is not either only increasing or only decreasing; rather, it is decreasing when $x < 0$ and increasing when $x > 0$. Therefore, we cannot apply the Transformation Theorem as is, but we can still apply the idea of its proof.

Note that
$$F_Y(y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

So, differentiating (to get the PDF), we find

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi y}}e^{-y/2}.$$

Note that this is known as the **Chi-squared** distribution.

# 6    Expected Value & Variance

Let $X$ be a continuous (potentially very complicated) random variable with PDF $f$. Suppose we were asked to pick *one real number* $\mu$, subject to a certain constraint, that best represents the random variable $X$. Of course, note that $X$ is a random variable and might be very different than just one single (deterministic) real number. That being said, we want to try our best to pick said real number that satisfies this constraint.

## 6.1    Expected Value

One natural thing to do is to pick a $\mu$ so that best represents the random variable (i.e., is closest to the random variable). So, we can consider the *squared distance*. In particular, we want to find a $\mu$ such that

$$(X - \mu)^2$$

is likely to be as close to 0 as possible. Note that $(X - \mu)^2$ is itself a random variable. Put it another way, we want to find a $\mu$ that minimize the random distance $(X - \mu)^2$. So, we want to minimize

$$\int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

By linearity, we have

$$\int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f(x) dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx.$$

Since $f(x)$ is a PDF, the third integral is just 1. Hence, we want to minimize

$$\int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \mu^2.$$

Differentiating with respect to $\mu$, we get

$$-2 \int_{\infty}^{\infty} x f(x) dx + 2\mu.$$

If we set this equal to 0 and then solve for $\mu$, we find the minimizer is

$$\mu = \int_{-\infty}^{\infty} x f(x) dx.$$

We note that $\mu$ is a **weighted average**. We integrate over all possible values of $x$, and then weigh them by their corresponding density $f(x)$.

---

**Definition 6.1: Expected Value**

Suppose that $X$ is a continuous random variable with PDF $f$. Then, the **expected value** (or **mean**) of $X$ is

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

---

**Remarks:**

- For some random variables, this expected value (integral) does not exist. We won't need to worry about this here, though.

- If $X$ is a **discrete** random variable with PMF $p(x) = \mathbb{P}(X = x)$, then $\mu = \mathbb{E}(X) = \sum_x x p(x)$.

The **mean** $\mu$ gives us valuable information about the "center" of a distribution. This, along with **standard deviation** $\mu$ (which tells us about the "spread" of a distribution), tells us the fundamental quantities in regards to the "shape" of a distribution.

(Example.) Suppose we roll a fair die, which takes values between 1 and 6, each with equal probability $\frac{1}{6}$. Then,

$$\mu = \sum_{k=1}^{6} k \cdot \frac{1}{6} = \frac{1}{6} \sum_{k=1}^{6} k = \frac{7}{2}.$$

Here, we're considering all possible values that the die can take (from $k = 1$ to 6). Each value has a probability of $\frac{1}{6}$.

While $\frac{7}{2}$ is not a value that the random variable can actually take (note that the random variable can only take values 1, 2, 3, 4, 5, and 6), but in some sense it is the expected value, or the average value. To see why this is the case, consider the following histogram:
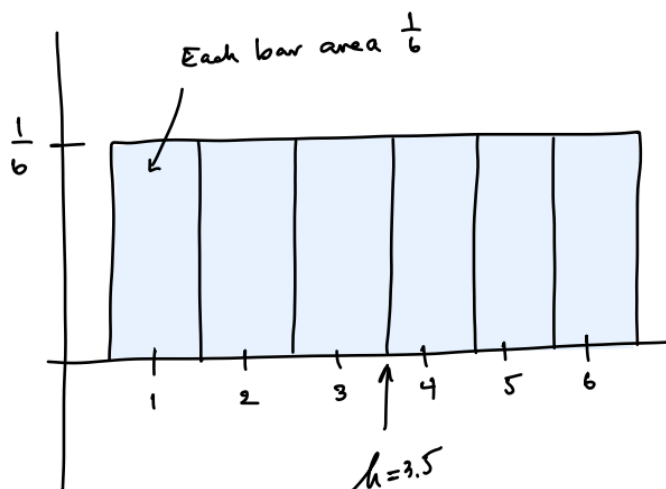


We can think of the 3.5 mark as the "center of mass."

## Theorem 6.1: Law of Unconscious Statistics (LotUS)

Suppose that $X$ is a random variable and $\phi : \mathbb{R} \mapsto \mathbb{R}$ is a function. Then,

1. If $X$ is continuous with PDF $f$, then

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

2. If $X$ is discrete with PMF $p$, then

$$\mathbb{E}[\phi(X)] = \sum_{x} \phi(x)p(x).$$

**Remarks:**

- It is enough *just* to know the PDF/PMF of $X$ to find the expected value of $Y = \phi(X)$.

- This shows us that expectation is linear. In particular, the **Linearity of Expectation** (LoE) says that if we let $X$ be a random variable, then we can suppose that $a, b \in \mathbb{R}$. Then, $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.

**Warning:**

- LoE does not imply that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

- Note that this is true if $X$ and $Y$ are independent, but not true otherwise.

## 6.2    Variance

> **Definition 6.2: Variance**
>
> The **variance** of a random variable $X$, denoted by $\sigma^2 = \text{Var}(X)$, is the expected squared distance of $X$ from its mean $\mu$. That is, $\sigma^2 = \mathbb{E}[(X - \mu)^2]$.

**Remarks:**

- This is telling us about the *spread* of the distribution; that is, how far away should we expect the random variable to be away from its mean.

- Recall that the mean $\mu$ is the real number that minimizes $\mathbb{E}[(X-a)^2]$ over all real numbers $a$. Whatever this quantity happens to be at the minimizer $a = \mu$ is called the variance $\sigma^2$.

> **Theorem 6.2**
>
> Let $X$ be a random variable and let $a, b \in \mathbb{R}$. Then,
>
> $$\text{Var}(aX + b) = a^2 \text{Var}(X) = a^2 \mathbb{E}[(X - \mu)^2].$$

**Remark:** Note that the variance of a constant is 0, since the distribution of a constant random variable has no "spread" at all.

**Warning:** While expected value is linear, variance is not. In other words, if $X$ and $Y$ are *independent*, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. However, this is not true in general.

Note that the formula $\sigma^2 = \mathbb{E}[(X - \mu)^2]$ is useful conceptually, but not so convenient for computations. Instead, we can use the Linearity of Expectation to obtain the following, more convenient, formula:

$$\sigma^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2) = \mathbb{E}(X^2) - 2\mu\mu + \mu^2 = \mathbb{E}(X^2) - \mu^2.$$

Therefore, the **computational variance formula** is given by

$$\boxed{\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2}.$$

The quantity $\mathbb{E}(X^2)$ is called the **second moment** of $X$, and can be computed using LotUS. That is, if $X$ is continuous, then

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx.$$

If $X$ is discrete, then

$$\mathbb{E}(X^2) = \sum_x x^2 p(x).$$

## 6.3    Standard Deviation

> **Definition 6.3: Standard Deviation**
>
> The square root of the variance, $\sigma = \sqrt{\text{Var}(X)}$, is called the **standard deviation** of $X$.

The reason for the name is that, for "most reasonable" distributions, the "bulk" of the mass/density deviates by at most $\sigma$ from the center $\mu$. That is, usually, "most" of the mass/density is on the values of

$$x \in [\mu - \sigma, \mu + \sigma].$$

> ### Theorem 6.3: Chebyshev's Inequality
>
> Let $X$ be a random variable with *both* mean $\mu$ and standard deviation $\sigma$. Then, for any real number $a > 0$, we have that
> $$\mathbb{P}(|X - \mu| \geq a\sigma) \leq \frac{1}{a^2}.$$

**Remarks:**

- In other words, for any distribution with a mean and variance, the probability that $X$ is at least $a = 2$ standard deviations from the center $\mu$ of the distribution is at most $\frac{1}{4}$.

- For many distributions, the actual probability is *much* smaller. However, this is an useful upper bound that works for all distributions.

This result follows by

> ### Theorem 6.4: Markov's Inequality
>
> Let $X$ be a non-negative random variable, i.e. $\mathbb{P}(X \geq 0) = 1$, with mean $\mu$ and $b > 0$ a positive number. Then, $\mathbb{P}(X \geq b) \leq \frac{\mu}{b}$.

---

*Proof.* (Continuous Case.) Since $X$ is non-negative, we know that

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{0}^{\infty} x f(x)\,dx.$$

Splitting the integral, we have

$$\int_{0}^{b} x f(x)\,dx + \int_{b}^{\infty} x f(x)\,dx \geq \int_{b}^{\infty} x f(x)\,dx \geq b \int_{b}^{\infty} f(x)\,dx = b\mathbb{P}(X \geq b).$$

Hence, $\mathbb{P}(X \geq b) \leq \frac{\mu}{b}$, as claimed. $\qquad\qquad\square$

---

## 6.4   Examples of Finding Expected Value and Variance

---

(Example.) If $X$ is Bernoulli$(p)$, then $\mu = p$ and $\sigma^2 = pq$, where $q = 1 - p$. Then,

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot q = p.$$

Note that $X^2$ has the same distribution as $X$, since $X$ only takes the values $0 = 0^2$ and $1 = 1^2$. Hence, $\mathbb{E}(X^2) = \mathbb{E}(X) = p$. Hence,
$$\mathrm{Var}(X) = p - p^2 = p(1 - p) = pq.$$

---

(Example.) If $X$ is Binomial$(n, p)$, then it is the sum of $n$ independent Bernoulli$(p)$ trials. By the Linearity of Expectation, we know that
$$\mathbb{E}(X) = np.$$

Since the trials are *independent*, we have that

$$\mathrm{Var}(X) = npq.$$

---

Note that this is the special case of the following fact.

> **Theorem 6.5**
>
> Suppose that $X_1, \ldots, X_n$ are IID with mean $\mu$ and variance $\sigma^2$. Then, their sum $S_n = \sum_{k=1}^{n} X_k$ has mean $\mathbb{E}(S_n) = n\mu$ and variance $\mathrm{Var}(S_n) = n\sigma^2$.

(Example.) If $X$ is Geometric($p$), then $\mathbb{E}(X) = \frac{1}{p}$. So, if $p$ is really small, then we should expect to wait a while before our first success; likewise, if $p$ is large, then we may not need to wait long before our first success. This is intuitive; in particular, the probability of success is $p$, so we should expect about 1 success in every $p$ trials. But, intuition aside, there are several ways to compute this.

- Approach 1.

  $$\mathbb{E}(X) = \sum_{k=1}^{\infty} kpq^{k-1} = p\sum_{k=1}^{\infty} kq^{k-1}.$$

  Recall that this is a *geometric* random variable, so we will use the geometric series; in particular,

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

  and

  $$\sum_{k=1}^{\infty} kq^{k-1} = \frac{d}{dq}\sum_{k=0}^{\infty} q^k.$$

  Hence,

  $$\mathbb{E}(X) = p\frac{d}{dq}\frac{1}{1-q} = p\frac{1}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}.$$

  Similarly, we can show that $\mathrm{Var}(X) = \frac{q}{p^2}$.

- Approach 2.

  Let $X$ be the number of trials until the first success. Then, we have

  $$\mathbb{E}(X) = 1p + (1 + \mathbb{E}(X))q.$$

  Solving for $\mathbb{E}(X)$ gives us the desired solution.

(Example.) A Poisson($\lambda$) has mean and variance $\mu = \sigma^2 = \lambda$. So,

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\infty}\frac{\lambda^k}{(k-1)!} = \lambda\sum_{k=0}^{\infty} e^{-\lambda}\underbrace{\frac{\lambda^k}{k!}}_{1} = \lambda.$$

Similarly, you can show that $\mathbb{E}(X^2) = \lambda(1 + \lambda)$ so that $\mathrm{Var}(X) = \lambda(1 + \lambda) - \lambda^2 = \lambda$.

(Example.) An Exponential($\lambda$) has $\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$. For this computation, the theorem following this example will be useful.

If $X$ is Exponential$(\lambda)$, then it is non-negative and $\mathbb{P}(X > x) = e^{-\lambda x}$. Hence,

$$\mathbb{E}(X) = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

---

### Theorem 6.6: Expectation Tail Sum for Non-Negative Random Variables

If $X$ is a non-negative random variable (i.e., $\mathbb{P}(X \geq 0) = 1$), then

1. If $X$ is discrete, then
$$\mathbb{E}(X) = \sum_{k=0}^\infty \mathbb{P}(X > k).$$

2. If $X$ is continuous, then
$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x) dx.$$

---

*Proof.* (Discrete.) Just note that

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^\infty kp(k) \\
&= p(1) + (p(2) + p(2)) + (p(3) + p(3) + p(3)) + \dots \\
&= (p(1) + p(2) + p(3) + \dots) + (p(2) + p(3) + \dots) + (p(3) + \dots) + \dots \\
&= \mathbb{P}(X > 0) + \mathbb{P}(X > 1) + \mathbb{P}(X > 2) + \dots
\end{aligned}$$

Hence, we're done.                                                                      $\square$

---

(Example.) If $X$ is Normal$(\mu, \sigma^2)$, then, indeed, $\mu$ is the mean and $\sigma^2$ is its variance. To see why this is the case, see lecture slides.

---

(Example.) If $X$ is Cauchy, then $\mathbb{E}(X)$ does not exist. Recall that a (standard) Cauchy has PDF

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

Note that

$$\int_{-\infty}^\infty \frac{x}{\pi(1 + x^2)} dx$$

diverges, since

$$\int_0^\infty \frac{x}{1 + x^2} dx = \infty.$$

---

## 6.5   Conditional Expectation

Recall that if $X$ is a discrete random variable with PMF $p$, and $B$ is an event with $\mathbb{P}(B) > 0$, then

$$p(x|B) = \frac{p(x)}{\mathbb{P}(B)}$$

is a probability distribution on $B$. This is the PMF of the random variable $X$, given $B$.

**Definition 6.4**

Let $X$ be a random variable with PMF $p$. Suppose that $\mathbb{P}(B) > 0$. Then, the conditional expectation of $X$ given $B$ is
$$\mathbb{E}(X|B) = \sum_x xp(x|B).$$

**Remark:** The situation is similar in the continuous case, but we instead have a conditional PDF

$$f(x|B) = \frac{f(x)}{\mathbb{P}(B)}$$

and the conditional expectation is given by

$$\mathbb{E}(X|B) = \int_{-\infty}^{\infty} xp(x|B)dx.$$

### 6.5.1 Law of Total Expectation

Just like how there is a Law of Total Probability, there is also a Law of Total Expectation.

**Theorem 6.7: Law of Total Expectation**

Let $X$ be a random variable on sample space $\Omega$. Suppose that $B_1, \ldots, B_n$ is a partition $\Omega$. Then,

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X|B_i)\mathbb{P}(B_i).$$

This is useful because, often, $\mathbb{E}(X)$ is sometimes difficult to find directly. However, if we condition on a well-chosen $B_i$, then it becomes manageable.

(Example.) In the gambling game "craps," a player makes a bet and then rolls a pair of dice. If the sum is 7 or 11, the player wins. If it is 2, 3, or 12, the player loses. If the sum is any other number $s$, the player continues to roll until either another $s$ (they win) or 7 (they lose) occurs (7 is lucky the first time). Now, let $R$ be the number of rolls in a single game of craps.

  1. Find $\mathbb{E}(R)$.

1. By the Law of Total Expectation, we have

$$\mathbb{E}(R) = \sum_{x=2}^{12} \mathbb{E}(R|X = x)\mathbb{P}(X = x),$$

where $X$ is the initial sum. Note that if

$$x \in \{2, 3, 7, 11, 12\},$$

then

$$\mathbb{E}(R|X = x) = 1$$

since the game is immediately over if we get one of those numbers. In particular,

- There is 1 way to get a 2 (11).
- There are 2 ways to get a 3 (12, 21).
- There are 6 ways to get a 7 (16, 61, 25, 52, 43, 34).
- There are 2 ways to get a 11 (56, 65).
- There is 1 way to get a 12 (66).

Hence,

$$\sum_{x \in \{2,3,7,11,12\}} \mathbb{E}(R|X = x)\mathbb{P}(X = x) = \frac{12}{36}.$$

Now, if

$$x \in \{4, 5, 6, 8, 9, 10\},$$

then we can use a similar argument to the one above. For example, when $x = 4$, we have 3 ways to get 4 (13, 31, 22). This gives us

$$\mathbb{P}(X = 4) = \frac{3}{36}.$$

There are also 6 ways to get 7 (16, 61, 25, 52, 43, 34). Therefore, the number of rolls $R$, given that the initial sum is $X = 4$, is distributed as $1 + G$, where $G$ is a geometric random variable (where the success is defined by rolling a 4 or 7) with success probability $p = \frac{9}{36}$. Note that the 1 is there because of the initial roll of the 4, but then we have to keep rolling. Thus, we get

$$\mathbb{E}(R|X = 4)\mathbb{P}(X = 4) = \left(1 + \frac{36}{9}\right)\frac{3}{36}.$$

To see how we got $\left(1 + \frac{36}{9}\right)$, recall that $E(X) = \frac{1}{p}$ if $X$ is Geometric($p$). So, by Linearity of Expectation, we get that the expected value of 1 (so it's 1) plus the expected value of the geometric (so it's 36/9).

So, by similar reasoning, we get

$$\mathbb{E}(R) = \frac{12}{36} + \left(1 + \frac{36}{9}\right)\frac{3}{36} + \left(1 + \frac{36}{10}\right)\frac{4}{36} + \left(1 + \frac{36}{11}\right)\frac{5}{36}$$

$$+ \left(1 + \frac{36}{11}\right)\frac{5}{36} + \left(1 + \frac{36}{10}\right)\frac{4}{36} + \left(1 + \frac{36}{9}\right)\frac{3}{36} = \frac{557}{165} \approx 3.375$$

So, on average, we expect a little bit more than 3 and 1/3 rolls of the dice in each of te game of craps.

(Example Problem.) Vito is lost in a maze. At the center of the maze, there are 3 paths. Path 1 leads out of the maze after a 2 minute walk. Paths 2 and 3 lead back to the center of the maze after 2 and 3 minute walks, respectively. Suppose that each time Vito is at the center of the maze he picks path $i$ with probability $i/6$. Show that, on average, Vito finds his way out in 15 minutes.

*Hint:* Use "First Step Analysis." That is, use the Law of Total Expectation, with respect to his first choice.

First, note that the probability that Vito picks path $i$, $P_i$, is given by

$$\mathbb{P}(P_1) = \frac{1}{6}, \quad \mathbb{P}(P_2) = \frac{2}{6} = \frac{1}{3}, \quad \mathbb{P}(P_3) = \frac{3}{6} = \frac{1}{2}.$$

Let $T$ be the time that Vito gets out of the maze. Then, we have

$$\mathbb{E}(T) = \mathbb{E}(T|P_1)\mathbb{P}(P_1) + \mathbb{E}(T|P_2)\mathbb{P}(P_2) + \mathbb{E}(T|P_3)\mathbb{P}(P_3).$$

Now, we note that

- $\mathbb{E}(T|P_1) = 2$ is the expected time by taking path 1.

- $\mathbb{E}(T|P_2) = 2 + \mathbb{E}(T)$ is the expected time by taking path 2. Note that we add $\mathbb{E}(T)$ because we end up back at the center.

- $\mathbb{E}(T|P_3) = 3 + \mathbb{E}(T)$ is the expected path by taking path 3. Again, note that we add $\mathbb{E}(T)$ because we end up back at the center.

Thus, we get

$$\mathbb{E}(T) = 2\frac{1}{6} + (2 + \mathbb{E}(T))\frac{1}{3} + (3 + \mathbb{E}(T))\frac{1}{2}$$

$$\implies \mathbb{E}(T) = \frac{1}{3} + \frac{2}{3} + \frac{1}{3}\mathbb{E}(T) + \frac{3}{2} + \frac{1}{2}\mathbb{E}(T)$$

$$\implies \mathbb{E}(T) - \frac{1}{3}\mathbb{E}(T) - \frac{1}{2}\mathbb{E}(T) = \frac{1}{3} + \frac{2}{3} + \frac{3}{2}$$

$$\implies \mathbb{E}(T)\left(1 - \frac{1}{3} - \frac{1}{2}\right) = \frac{1}{3} + \frac{2}{3} + \frac{3}{2}$$

$$\implies \mathbb{E}(T) = \frac{\frac{1}{3} + \frac{2}{3} + \frac{3}{2}}{1 - \frac{1}{3} - \frac{1}{2}}$$

$$\implies \mathbb{E}(T) = 15,$$

as expected.

### 6.5.2   Martingales

Informally, we can think of martingales as random processes that encapsulate the idea of a *fair game*.

> **Definition 6.5**
>
> Let $(X_0, X_1, X_2, \dots)$ be a sequence of random variables and $\phi$ a function. Let $M_n = \phi(X_n)$. The sequence $(M_0, M_1, M_2, \dots)$ is called a **martingale** (MG) with respect to $(X_0, X_1, X_2, \dots)$ if, for any $n$ and any $x_0, x_1, \dots, x_n$, we have that
>
> $$\mathbb{E}(M_{n+1} - M_n | X_n = x_n, \dots, X_0 = x_0) = 0.$$

**Remark:** If we think of $M_n$ as the total winnings after $n$ bets by a gambler, then $(M_0, M_1, M_2, \dots)$ is a "fair game" in the sense that neither the gambler nor the "house" has an advantage. In other words, after

the $(n+1)$th game, we would expect no additional gain (hence why this is equal to 0).

> **Theorem 6.8**
>
> For any random variables $X$ and $Y$, we have that
>
> $$\mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}(X).$$

**Remarks:**

- Recall that $\mathbb{E}(M_{n+1}|X_n,\ldots,X_0) = M_n$ for a MG. Then, taking expectations on both sides, it follows that $\mathbb{E}(M_{n+1}) = \mathbb{E}(M_n)$ for all $n$. Hence, for all (deterministic) times $n \geq 0$, we have that $\mathbb{E}(M_n) = \mathbb{E}(M_0)$.

- These are a powerful tool, which can lead to quick or slick proofs of things that would be computationally challenging otherwise.

> (Example.) Recall, from Lecture 1, that Peter and Paul keeps flipping a coin. If it is "Heads," then Peter wins \$1. Otherwise, Peter loses \$1. Let $X_i$ be Peter's winning after the $i$th bet. Note that $X_0 = 0$.
>
> To see that this is a MG, note that the series of flips are independent. Hence, for any $n$ and any $x_1,\ldots,x_n$, we have
>
> $$\mathbb{E}(X_{n+1} - X_n|X_n = x_n,\ldots,X_1 = x_1) = (1)\frac{1}{2} + (-1)\frac{1}{2} = 0.$$

Recall that $\mathbb{E}(M_n) = \mathbb{E}(M_0)$ for all (deterministic) times $n \geq 0$. However, this is not necessarily true for *random $T$*. Hence, we restrict to a special class of random times.

> **Definition 6.6**
>
> A time $T$ is a **stopping time** (ST) if and only if, for any $n$, to know if $T = n$, we only need to know the values of $X_0, X_1, \ldots, X_n$. That is, we do not need any information about the future after time $n$.

For example, the first time we visit 0 is a ST, but the last time we visit 0 is not (in order for us to know if time $n$ is 0, we need to know the full history).

> **Theorem 6.9: The Optional Stopping Theorem (OST)**
>
> If $(M_0, M_1, \ldots)$ is a MG and $T$ is a ST, then $\mathbb{E}(M_T) = \mathbb{E}(M_0)$ if the following conditions are satisfied:
>
> 1. $M_n$ is bounded until time $T$, and
>
> 2. $\mathbb{P}(T < \infty) = 1$.

> (Example: Gambler's Ruin). Suppose that Peter currently has \$1. Furthermore, suppose that they play instead with a coin that comes up Heads with probability $p \neq 1/2$. What is the probability $\mathbb{P}(J)$ that Peter wins the "Jackpot" (\$N) before going "bust" (\$0)?

For this biased RW, we know that
$$M_n = (q/p)^{X_n}$$
is a MG, where $X_n$ is Peter's winnings. Note that here
$$\phi(x) = (q/p)^x.$$

Then,
$$\mathbb{E}(M_{n+1} - M_n | X_n = x_n, \ldots, X_1 = x_1) = (q/p)^{x_n}(q/p - 1)p + (q/p)^{x_n}(p/q - 1)q$$
$$= (q/p)^{x_n}[(q - p) + (p - q)]$$
$$= 0.$$

Now, let $T$ be the first time that $X_n \in \{0, N\}$. By the OST, we know that
$$q/p = \mathbb{E}(M_T) = 1 \cdot \mathbb{P}(J^C) + (q/p)^N \cdot \mathbb{P}(J).$$

Hence,
$$\mathbb{P}(J) = \frac{(q/p) - q}{(q/p)^N - 1}.$$