Probability Proof Summary

Ari Feiglin

July, 2022

Contents

0.1 Prerequisites

Definition 0.1.1:

In order to reduce any ambiguity, I will define the following substitute for the set of naturals:
Given $m \in \mathbb{Z}$:

$$\mathbb{N}_m := \{n \in \mathbb{Z} \mid n \geq m\}$$

Definition 0.1.2:

Since it pops up a lot in combinatorics and probability, it is useful to define:

$$[n] := \{m \in \mathbb{N}_1 \mid m \leq n\}$$

Definition 0.1.3:

I define $\mathcal{P}_k(A)$ to be the set of all k-length subsets of A:

$$\mathcal{P}_k(A) := \{B \subseteq A \mid |B| = k\}$$

Suppose the cardinality of A is n. I'll define the cardinality of $\mathcal{P}_k(A)$ to be:

$$\binom{n}{k} := |\mathcal{P}_k(A)|$$

This is called a **binomial coefficient**.

Definition 0.1.4:

The **symmetric group** of a set A, denoted S_A, is the set of all bijections from A to itself.
S_n is another and more common way of writing $S_{[n]}$.

1 Combinatorics

1.1 Orderings

Proposition 1.1.1:

There are $\frac{n!}{(n-k)!}$ injective functions from $[k]$ to $[n]$.

Proof:

Let A be this set of injective functions.
Let's create a bijection:

$$f : [n] \times [n-1] \times \cdots \times [n-k+1] \longrightarrow A$$

We know that forall $a_1, \ldots, a_k \in [n]$ there exists a bijection:

$$f_{\{a_1,\ldots,a_k\}} : [n-k] \longrightarrow [n] \setminus \{a_1, \ldots, a_k\}$$

So we can define f like so:

$$f(a_1, a_2, \ldots, a_{n-k}) = g$$

Where:

$$g(1) = a_1$$

And:

$$g(i) = f_{\{g(1),\ldots,g(i-1)\}}(a_i)$$

For $i \geq 2$.
This is well-defined and g is an injection since $g(i) \in [n] \setminus \{g(1), \ldots, g(i-1)\}$ (this can be shown inductively. It is not trivial since we must prove that $g(1) \neq \cdots g(i-1)$ in order to show that $f_{\{g(1),\ldots,g(i-1)\}}$ is the intended bijection.)
f is injective since if:

$$f(a_1, \ldots, a_{n-k})g_1 = g_2 = f(b_1, \ldots, b_{n-k})$$

Then:

$$g_1(1) = g_2(1) \implies a_1 = b_1$$

And:

$$g_1(2) = g_2(2) \implies f_{\{a_1\}}(a_2) = f_{\{a_1\}}(b_2) \implies a_2 = b_2$$

Since $f_{\{a_1\}}$ is a bijection.
And so on.
Now we must prove that f is surjective. Suppose g is an injection. Notice then that:

$$f(g(1), f^{-1}_{\{g(1)\}}(g(2)), f^{-1}_{\{g(1),g(2)\}}(g(3)), \ldots, f^{-1}_{\{g(1),\ldots,g(k-1)\}}(g(k))) = g$$

Let h be the left side ($f(\ldots)$), then:

$$h(i) = f_{\{g(1),\ldots,g(i-1)\}}\left(f^{-1}_{\{g(1),\ldots,g(i-1)\}}(g(i))\right) = g(i)$$

So $h = g$ as required.
Therefore f is a bijection and:

$$|A| = n \cdot (n-1) \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

As required.

■

Note:

This proof is just a more rigorous wording of the classic construction of the injective functions.
To construct an injective function, first choose the image of 1. There are n choices for this. Then there are $n-1$ choices

for 2, and so on. In the end we get that in total there are:

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

choices and thus injective functions.

Lemma 1.1.2:

If A and B are finite sets with the same cardinality, then every injection from A to B is a bijection.

Proof:

Suppose $f : A \longrightarrow B$ is an injection. This means that $|\text{Im} f| = |A| = |B|$.
Suppose, for the sake of a contradiction, that f is not a surjection. Then there exists a $b \in B$ which has not origin in A, that is $b \notin \text{Im} f$.
But we know that $\text{Im} f \sqcup \{b\} \subseteq B$, so:

$$\begin{aligned} & |\text{Im} f \sqcup \{b\}| \leq |B| \\ \implies & |\text{Im} f| + 1 \leq |B| \\ \implies & |B| + 1 \leq |B| \\ \implies & 1 \leq 0 \end{aligned}$$

In contradiction. So f is a surjection, and therefore a bijection, as required.

■

Theorem 1.1.3:

There are $n!$ permutations of a set of cardinality n. That is, $|S_n| = n!$.

Proof:

We know that every injection from $[n]$ to $[n]$ is a bijection, and vice versa. So S_n is the set of injections from $[n]$ to $[n]$, of which we proved there are:

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

As required.

■

Lemma 1.1.4:

If P is a partition of A such that every equivalence class has equal cardinality:

$$|A| = |P| \cdot p$$

Where p is the cardinality of an equivalence class.

Proof:

Suppose $[a]$ is an equivalence class of A. Then we know that for every equivalence class $[b]$, there exists a bijection:

$$f_{[b]} : [a] \longrightarrow [b]$$

Since they have equal cardinalities.
We'll define a function:

$$f : P \times [a] \longrightarrow A$$

Where:

$$f([b], \alpha) = f_{[b]}(\alpha)$$

This is injective since:

$$f([b], \alpha) = f([c], \beta) \implies f_{[b]}(\alpha) = f_{[c]}(\beta)$$

Since the codomain of $f_{[x]}$ is $[x]$, and if $[x] \neq [y]$ then $[x] \cap [y] = \varnothing$ as P is a partition, this means that $[b] = [c]$. And since $f_{[b]}$ is a bijection, this means that $\alpha = \beta$.
So $([b], \alpha) = ([c], \beta)$, which means f is injective.
Now, suppose $b \in A$, then:

$$f([b], f_{[b]}^{-1}(b)) = f_{[b]}\left(f_{[b]}^{-1}(b)\right) = b$$

So f is surjective.
This means that f is a bijection.

■

Proposition:

There are $(n-1)!$ distinct ways to place n people around a circular table.

Proof:

This is the same as asking how many distinct permutations there are if we define the equivalence class of a permutation $\sigma \in S_n$ by:

$$[\sigma] = \{\tau \in S_n \mid \exists i \in \mathbb{Z} : \tau(x) = \sigma(x + i \bmod n)\}$$

As $x + i \bmod n$ corresponds to a shift of i spots about the table.
We know for all $i \in \mathbb{Z}$, $x + i \bmod n = x + (i \bmod n) \bmod n$, and $0 \leq i \bmod n < n$. So:

$$[\sigma] = \{\tau \in S_n \mid \exists 0 \leq i < n : \tau(x) = \sigma(x + i \bmod n)\}$$

And for every $0 \leq i \neq j < n$:

$$x + i \bmod n \neq x + j \bmod n \implies \sigma(x + i \bmod n) \neq \sigma(x + j \bmod n)$$

Therefore:

$$|[\sigma]| = n$$

By **lemma 1.1.4**, this means that the number of distinct permutation is:

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

As required.

■

Theorem 1.1.5:

The number of distinct ways to order k black balls and $n-k$ white ones is:

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

Proof:

Let the balls form the series $\{a_i\}_{i=1}^{n}$ where a_i is 1 (black) for $i \leq k$ and 0 (white) for all other a_is. We define the equivalence class between permutations of the n balls:

$$[\sigma] = \left\{\tau \in S_n \;\middle|\; a_{\tau(i)} = a_{\sigma(i)}\right\}$$

This represents all the permutations of the balls which give the same ordering as σ. We want to count the number of distinct permutations there are, which is the number of distinct equivalence classes.
Let:

$$A_\sigma := \left\{n \geq i \in \mathbb{N}_1 \;\middle|\; a_{\sigma(i)} = 1\right\}$$

And:

$$B_\sigma := [n] \setminus A_\sigma$$

We will define a bijection:

$$f : S_{A_\sigma} \times S_{B_\sigma} \longrightarrow [\sigma]$$

Where:

$$f(\sigma_A, \sigma_B) = \tau$$

Where τ is defined by:

$$\tau(x) = \begin{cases} \sigma_A(x) & x \in A_\sigma \\ \sigma_B(x) & x \in B_\sigma \end{cases}$$

We need to show that this is well-defined. Firstly, this is a bijection because A_σ and B_σ are disjoint and σ_A and σ_B are bijections. Suppose $i \in [n]$, if $i \in A_\sigma$ then by definition $a_{\sigma(i)} = 1$, and $a_{\tau(i)} = a_{\sigma_A(i)}$, but $\sigma_A(i) \in A_\sigma$, so $a_{\sigma_A(i)} = 1$, as required. The proof is similar if $i \in B_\sigma$.
Now, let's show that f is injective. Suppose:

$$f(\sigma_A, \sigma_B) = \tau_1 = \tau_2 = f(\pi_A, \pi_B)$$

This means that for every $i \in A_\sigma$: $\tau_1(i) = \tau_2(i)$, but we know that this just means $\sigma_A(i) = \pi_A(i)$, therefore $\sigma_A = \pi_A$. Similar for B. Since the two sets are finite, this means that f is a bijection.
And we know that the cardinality of A_σ is k (since σ is a bijection and there are k a_is which equal 1), and therefore B_σ has a cardinality of $n-k$. So:

$$|\sigma| = k!\cdot(n-k)!$$

So by **lemma 1.1.4**, this means the number of equivalence classes is:

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

■

Corollary 1.1.6:

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

Proof:

Let $\{a_i\}_{i=1}^n$ be defined similarly to as it was above. Let P the set of distinct orderings of this set. As per the theorem above, $|P| = \frac{n!}{k!\cdot(n-k)!}$. We will construct a bijection:

$$f : P \longrightarrow \mathcal{P}_k([n])$$

We define it like so:

$$f([\sigma]) = \left\{ m \in [n] \;\middle|\; a_{\sigma(m)} = 1 \right\}$$

We know this must have a cardinality of k since there are k a_is which are equal to 1, and permutations in the same equivalence class map the same a_is to 1, so the function is well-defined.
Let us prove that f is injective. If $f(\sigma) = f(\tau)$ then

$$a_{\sigma(m)} = 1 \iff a_{\tau(m)} = 1$$

Which means that $a_{\sigma(m)} = a_{\tau(m)}$, and therefore $[\tau] = [\sigma]$, as required.
Now suppose $S = \{x_1, \ldots, x_k\} \in \mathcal{P}_k([n])$. We can define a permutation like so:

$$\sigma(x_i) = i$$

And since $|[n] \setminus S| = n - k$, there is a bijection between $[n] \setminus S$ and $\{k+1, \ldots, n\}$. So we can map the values in $[n] \setminus S$ (non-x_i values) bijectively to $\{k+1, \ldots, n\}$.
This defines a bijection σ in S_n. Now notice that since $a_i = 1$ only for $i \leq k$:

$$\left\{ m \in [n] \;\middle|\; a_{\sigma(m)} = 1 \right\} = \{ m \in [n] \mid \sigma(m) \leq k \} = \{x_1, \ldots, x_k\} = S$$

So:

$$f([\sigma]) = S$$

And f is therefore a surjection.
Therefore f is bijective and:

$$\binom{n}{k} = |\mathcal{P}_k([n])| = |P| = \frac{n!}{k! \cdot (n-k)!}$$

As required.

■

Proposition (Pascal's Identity):

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

This identity is named after **Blaise Pascal**, one of the pioneers of early probability theory.

Proof:

> **Note:**
>
> This can be proved algebraically, by applying the formula for binomial coefficient which we proved earlier. But this proof is dry and does not reveal much about the inner nature of the identity.

This is a simple proof. All it requires is a construction of a bijection:

$$f : \mathcal{P}_k([n]) \longrightarrow \mathcal{P}_{n-k}([n])$$

The construction is quite simple and natural:

$$f(S) = [n] \setminus S$$

This is obviously well-defined, and is a bijection since it is its own inverse.

■

Proposition 1.1.7 (Pascal's Rule):

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

Proof:

A good look at the proposition reveals what the theorem really means: there is a partition of $\mathcal{P}_k([n])$ into two sets isomorphic to $\mathcal{P}_k([n-1])$ and $\mathcal{P}_{k-1}([n-1])$ respectively. So let's attempt to find (one) of these partitions.
We can define the partition into two sets:

$$A = \{S \in \mathcal{P}_k([n]) \mid n \in S\}$$

And:

$$B = \mathcal{P}_k([n]) \setminus A = \{S \in \mathcal{P}_k([n]) \mid n \notin S\}$$

First, let's show that A is isomorphic to $\mathcal{P}_{k-1}([n-1])$. Let $S \in A$. We can map it to $S \setminus \{n\}$. This is well-defined since it has a cardinality of $|S| - 1 = k - 1$ and is a subset of $[n] \setminus \{n\} = [n-1]$. Since for every $S \in A$, $n \in S$, this is injective (since $S = S \setminus \{n\} \cup \{n\}$). It is surjective since given $S' \in \mathcal{P}_{k-1}([n-1])$, $S' \cup \{n\}$ will be mapped to it.
And B is actually just equal to $\mathcal{P}_k([n-1])$. This is because if $S \in B$ then $S \subseteq [n] \setminus \{n\} = [n-1]$, and S has a cardinality of k.
And $A \sqcup B = \mathcal{P}_k([n])$ trivially.
Therefore:

$$|A| + |B| = |\mathcal{P}_k([n])| \implies \binom{n-1}{k-1} + \binom{n-1}{k} = \binom{n}{k}$$

As required.

■

Definition 1.1.8:

A **multiset** in a universe $\mathbb{U}$ is a pair:

$$A = (\mathbb{U}, m)$$

Where $\mathbb{U}$ is a set and m is a function:

$$m\colon \mathbb{U} \longrightarrow \mathbb{N}_0$$

This is called A's **multiplicity function** and is also denoted m_A.
Multisets are a way of representing sets where elements can be repeated a finite number of times. Basically, for every a in A, $m(a)$ represents how many times a is in the multiset.
Let $\mathcal{M}(\mathbb{U})$ be the set of all multisets in the bU $\mathbb{U}$, that is:

$$\mathcal{M}(\mathbb{U}) \coloneqq \{(\mathbb{U}, m) \mid m\colon \mathbb{U} \longrightarrow \mathbb{N}_0\} = \mathbb{U} \times (\mathbb{N}_0)^{\mathbb{U}}$$

Given a multiset A in a universe $\mathbb{U}$, we define its **support** to be the set of all elements in A:

$$\operatorname{supp}(A) \coloneqq \{a \in \mathbb{U} \mid m(a) > 0\}$$

And we say that $a \in A$ if $a \in \operatorname{supp}(A)$.
We define the **cardinality** of a finite multiset A (a multiset is finite if its support is):

$$|A| \coloneqq \sum_{a \in \operatorname{supp} A} m(a)$$

Finally, we define the set of all k-length multisets over $\mathbb{U}$ as:

$$\mathcal{M}_k(\mathbb{U}) = \{A \in \mathcal{M}(\mathbb{U}) \mid |A| = k\}$$

And we define its cardinality to be:

$$\left(\!\binom{|\mathbb{U}|}{k}\!\right) \coloneqq |\mathcal{M}_k(\mathbb{U})|$$

This is called the **multiset coefficient**.

Proposition 1.1.9:

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

Proof:

So we need to find the cardinality of $\mathcal{M}_k([n])$.
Firstly, recall that by **corollary 1.1.6** the number of distinct orderings of k of one object and $n-1$ of another is $\binom{n+k-1}{k}$.
So let's create a bijection from the set of distinct orderings of k one object and $n-1$ of another.
These orderings can be uniquely characterized by the indexes of the $n-1$ objects, which are the series $\{p_i\}_{i=1}^{n-1}$ where $1 \le p_1 < \cdots < p_{n-1} \le n+k-1$. For simplicity, let's define $p_0 \coloneqq 0$ and $p_n \coloneqq n+k$.
So now let's define the bijection:

$$f\left(\{p_i\}_{i=0}^{n}\right) = ([n], m)$$

Where for every $k \in [n]$:

$$m(k) = p_k - p_{k-1} - 1$$

Let's prove that this is well-defined. Firstly, $p_k > p_{k-1} \implies p_k - p_{k-1} - 1 \ge 0$, so $m(k) \in \mathbb{N}_0$ as required.
Secondly:

$$\sum_{k=1}^{n} m(k) = \sum_{k=1}^{n} p_k - p_{k-1} - 1 = \sum_{k=1}^{n} (p_k - p_{k-1}) - n$$

This is a telescopic sum, which is equal to:

$$= p_n - p_0 - n = n + k - n = k$$

Which means this multiset has a cardinality of k, as required.
Now, let's show that f is an injection.

Suppose:

$$f\left(\{p_i\}_{i=0}^{n}\right) = f\left(\{q_i\}_{i=0}^{n}\right)$$

That means for every $k \in [n]$:

$$p_k - p_{k-1} - 1 = q_k - q_{k-1} - 1 \implies p_k - p_{k-1} = q_k - q_{k-1}$$

And through simple induction it can be shown that $p_k = q_k$ (recall that $p_0 = q_0 = 0$ for the base).
So f is injective.
Now, suppose $([n], m) \in \mathcal{M}_k([n])$ is a multiset, we can define:

$$p_k \coloneqq \sum_{i=1}^{k} (m(i)) + k$$

Which is a valid indexing since $p_k < p_{k+1}$, $p_0 = 0$, and $p_n = \sum_{i=1}^{k} m(i) + n = k + n$.
And notice that:

$$p_k - p_{k-1} = m(k) + 1 \implies m(k) = p_k - p_{k-1} - 1$$

Which means that:

$$f\left(\{p_i\}_{i=0}^{n}\right) = ([n], m)$$

So every multiset has an origin, therefore f is surjective.
So f is a bijection, which means that:

$$\left(\!\binom{n}{k}\!\right) = |\mathcal{M}_k([n])| = \binom{n+k-1}{k}$$

As required.

■

Proposition 1.1.10:

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

Proof:

As shown earlier, $\mathcal{M}_k([n])$ is isomorphic to the set of orderings of $n-1$ of one object and k of another. So for instance, it is isomorphic to the set of orderings of $n-1$ black balls and k white balls.
But we can flip which object is which, so this is isomorphic to the orderings of k black balls and $n-1$ white balls.
And this is isomorphic by the (proof of the) previous proposition to $\mathcal{M}_{n-1}([k+1])$.
So all in all $\mathcal{M}_k([n])$ is isomorphic to $\mathcal{M}_{n-1}([k+1])$, so:

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

As required.

■

Proposition 1.1.11:

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

Proof:

We will create a partition of $\mathcal{M}_k([n])$ into subsets which have the required cardinality. The partition which accomplishes this is defined as follows:

$$\mathcal{M}_k([n]) = A \sqcup B$$

Where:

$$A \coloneqq \{M \in \mathcal{M}_k([n]) \mid n \notin M\} \quad B \coloneqq \{M \in \mathcal{M}_k([n]) \mid n \in M\}$$

Notice that $A = \mathcal{M}_k([n-1])$.
We can define a bijection f from B to $\mathcal{M}_{k-1}([n])$ as follows:

$$f(M) = \tilde{M}$$

Where:

$$m_{\tilde{M}}(i) = \begin{cases} m_M(i) & i \neq n \\ m_M(i) - 1 & i = n \end{cases}$$

This is well-defined because the sum of $m_{\tilde{M}}$ is one less than the sum of m_M, which is k. So the sum is $k-1$, as required.
It is obviously injective, since we have all the necessary knowledge about m in its image.
It is also obviously surjective, since given $m_{\tilde{M}}$, we can define m like so:

$$m(i) = \begin{cases} m_{\tilde{M}}(i) & i \neq n \\ m_{\tilde{M}}(i) + q & i = n \end{cases}$$

So f is a bijection, therefore:

$$|B| = \left(\!\!\binom{n}{k-1}\!\!\right)$$

So all in all we know:

$$|\mathcal{M}_k([n])| = |A| + |B| \implies \left(\!\!\binom{n}{k}\!\!\right) = \left(\!\!\binom{n-1}{k}\!\!\right) + \left(\!\!\binom{n}{k-1}\!\!\right)$$

As required.

■

Theorem 1.1.12:

Given a set A with n elements, the number of ways to choosen k elements from A is:

- If order matters (so choosing a then b is different from choosing b then a) and repition is allowed (we can choose a twice for example):
$$n^k$$
- If order matters, but repetition is not allowed:
$$\frac{n!}{(n-k)!}$$
- If order doesn't matter, but repetition is allowed:
$$\left(\!\binom{n}{k}\!\right) = \binom{n+k-1}{k}$$
- If order doesn't matter, and repetition is not allowed:
$$\binom{n}{k}$$

This should give insight as to why what we've been discussing is significant.

Proof:

- If order matters and repetition is allowed, then this is just analogous to the number of tuples over A with length k, this is the set A^k, which has a cardinality of n^k.
- If order matters, but repetition is not allowed, then this is just analogous to the number of tuples over A with length k, but all elements in the tuples are distinct. And tuples are analogous to functions, so this is analogous to the injective functions from $[k]$ to A, of which there are $\frac{n!}{(n-k)!}$.
- If order doesn't matter, and repetition is allowed, then we're just choosing a multiset of length k from A. And we know there are $\left(\!\binom{n}{k}\!\right) = \binom{n+k-1}{k}$ of those.
- If order doesn't matter and repetition is not allowed, then we're just choosing a set of cardinality k from A, in other words a k-length subset of A. And we know there are $\binom{n}{k}$ of those.

■

1.2 Multinomials

Definition 1.2.1:

Given a tuple I, $[I]_i$ is the element in the i-th position of the tuple.
If I has a size of n, then for a $\sigma \in S_n$, I define $\sigma(I)$ as the tuple where:

$$[\sigma(I)]_i = [I]_{\sigma(i)}$$

Lemma 1.2.2:

$$[(\sigma \circ \tau)(I)]_i = [\sigma(I)]_{\tau(i)}$$

Proof:

Notice that:

$$[\sigma(I)]_{\tau(i)} = [I]_{\sigma\circ\tau(i)}$$

And:

$$[(\sigma \circ \tau)(I)] = [I]_{\sigma\circ\tau(i)}$$

As required.

■

Lemma 1.2.3:

Suppose $S = \{s_1, \ldots, s_\ell\}$.
The number of tuples of length n over S which have k_i occurrences of s_i is:

$$\frac{n!}{k_1! \cdots k_\ell!}$$

(Assuming $\sum_{i=1}^{\ell} k_i = n$)

Proof:

Let:

$$I = \left(\underbrace{s_1, \ldots, s_1}_{k_1 \text{ times}}, \ldots, \underbrace{s_\ell, \ldots, s_\ell}_{k_\ell \text{ times}} \right)$$

All other tuples which satisfy the criteria are permutations of this one.
We define equivalence the equivalence class of $\sigma \in S_n$ as the set of all permutations which produce the same tuple:

$$[\sigma] = \{\tau \in S_n \mid \sigma(I) = \tau(I)\}$$

We need to find the number of distinct equivalence classes there are. To do so we will find the cardinality of each equivalence class.
Let:

$$\sigma_j = \left\{ i \in [n] \,\middle|\, [\sigma(I)]_i = s_j \right\}$$

For $j \in [\ell]$. This is the set of all indexes in the permuted tuple of the elements s_j (notice that $[\tau] = [\sigma]$ if and only if $\sigma_j = \tau_j$ for evey j). It is important to note that $\{\sigma_j\}_{j=1}^{\ell}$ partitions $[n]$. If $i \in [n]$ then $[\sigma(I)]_i = s_j$ for some j (which is distinct since the s_js are distinct), and therefore $i \in \sigma_j$ for only that j.

Since there are k_j indexes which equal s_j, $\left|\sigma_j\right| = k_j$.
We will construct a bijection:

$$f\colon S_{\sigma_1} \times \cdots \times S_{\sigma_\ell} \longrightarrow [\sigma]$$

by:

$$f(\pi_1, \ldots, \pi_\ell) = \tau$$

Where τ is defined by:

$$\tau = \sigma \circ \pi$$

If $i \in [n]$ then there exists a unique $j \in [\ell]$ such that $i \in \sigma_j$. Then we define:

$$\pi(i) = \pi_j(i)$$

π is a bijection since π_j is, and therefore so is τ. Notice that if $i \in \sigma_j$, then $[(\sigma \circ \pi)(I)]_i = [\sigma(I)]_{\pi(i)}$, and since $i \in \sigma_j$, then $\pi(i) = \pi_j(i) \in \sigma_j$, so

$$[(\sigma \circ \pi)(I)]_i = [\tau(I)]_i = s_j = [\sigma(I)]_i$$

Which means that $\tau \in [\sigma]$, so f is well-defined.
This function corresponds to taking $\sigma(I)$ and permuting its indexes as to not mess up their elements.
This is injective since if:

$$f(\pi_1, \ldots, \pi_\ell) = f(\pi'_1, \ldots, \pi'_\ell)$$

Then:

$$\sigma \circ \pi = \sigma \circ \pi' \implies \pi = \pi'$$

But that means that $\pi(i) = \pi'(i)$ for every i, and therefore for every j:

$$\pi_j(i) = \pi'_j(i)$$

(For relevant is), and therefore $\pi_j = \pi'_j$. So f is injective.
This is surjective. If $\tau \in [\sigma]$, then we need to find π_js such that their corresponding π is equal to $\sigma^{-1} \circ \tau$.
Suppose $i \in \sigma_j$, we define:

$$\pi_j(i) = \sigma^{-1} \circ \tau(i)$$

We need to show that this is well-defined. So we must show that $\sigma^{-1} \circ \tau(i) \in \sigma_j$.

$$[\sigma(I)]_{\sigma^{-1} \circ \tau(i)} = [I]_{\sigma \circ \sigma^{-1} \circ \tau(i)} = [I]_{\tau(i)} = [\tau(I)]_i$$

And since $[\tau] = [\sigma]$, this is equal to $[\sigma(I)]_i = s_j$. So:

$$[\sigma(I)]_{\sigma^{-1} \circ \tau(i)} = s_j \implies \sigma^{-1} \circ \tau(i) \in \sigma_j$$

As required.
And since for all relevant i: $\pi_j(i) = \sigma^{-1} \circ \tau(i)$, this means $\pi = \sigma^{-1} \circ \tau$, as required.
So f is a bijection. This means:

$$|[\sigma]| = \left|S_{\sigma_1}\right| \cdots \left|S_{\sigma_\ell}\right| = k_1! \cdots k_\ell!$$

By **lemma 1.1.4**, this the number of distinct permutations is:

$$\frac{n!}{k_1! \cdots k_\ell!}$$

As required.

■

Definition 1.2.4:

We define the result of the previous lemma to be something called a **multinomial coefficient**:

$$\binom{n}{k_1,\dots,k_\ell} = \frac{n!}{k_1!\cdots k_\ell!}$$

If $\sum_{i=1}^{\ell} k_i = n$.

A natural question to ask at this point is why is this called a **multinomial** coefficient: What is its relation with binomial coefficients?

Proposition 1.2.5:

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

(The left side is a **binomial coefficient** while the right side is a **multinomial coefficient**.)

Proof:

Notice that $\binom{n}{k,n-k}$ represents the number of distinct tuples with k of one element and $n-k$ of another. This is analogous, equivalent actually, to the number of ways to order k of one element and $n-k$ of another. And as we discusses in the proof of **lemma 1.1.6**, this is equal to the cardinality of $\mathcal{P}_k([n])$.
Therefore

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

As required.

■

Definition 1.2.6:

Given a tuple I of length n over a set of elements S, I define for every $s \in S$:

$$\#_s I := \left|\{i \in [n] \;\middle|\; [I]_i = s\}\right|$$

Which is the number of occurrences of s in the tuple I.

Theorem 1.2.7 (The Multinomial Theorem):

$$\left(\sum_{i=1}^{\ell} a_i\right)^n = \sum_{k_1+\cdots+k_\ell=n} \binom{n}{k_1,\ldots,k_\ell} \cdot a_1^{k_1} \cdots a_\ell^{k_\ell}$$

This is a generalization of the more famous **Binomial Theorem**, which we will discuss soon.

Proof:

We know:

$$\left(\sum_{i=1}^{\ell} a_i\right)^n = \sum_{i_1=1}^{\ell} a_{i_1} \cdots \sum_{i_n=1}^{\ell} a_{i_n} = \sum_{i_1=1}^{\ell} \cdots \sum_{i_n=1}^{\ell} \left(a_{i_1} \cdots a_{i_n}\right)$$

Let I be defined as the tuple:

$$I := (i_1, \ldots, i_n)$$

We can rewrite the sum as:

$$\sum_{I \in [\ell]^n} \prod_{i \in I} a_i$$

We can partition $[\ell]^n$ like so: for every $\{k_i\}_{i=1}^{\ell}$ such that $\sum_{i=1}^{\ell} k_i$, create a subset of $[\ell]^n$ defined as follows:

$$\{I \in [\ell]^n \mid \forall j \in [\ell] : \#_j I = k_j\}$$

These sets are obviously disjoint, and their union is $[\ell]^n$ since for every $I \in [\ell]^n$ define $k_j = \#_j I$.
Suppose S is one of these sets characterized by $\{k_i\}_{i=1}^{\ell}$. Then for every $I \in S$:

$$\prod_{i \in I} a_i = a_1^{k_1} \cdots a_\ell^{k_\ell}$$

Since there are k_i instances of i in I.
And we know by the previous lemma that $|S| = \binom{n}{k_1,\ldots,k_\ell}$. So summing over S gives:

$$\binom{n}{k_1,\ldots,k_\ell} a_1^{k_1} \cdots a_\ell^{k_\ell}$$

And since S is characterized by and only by $\{k_i\}_{i=1}^{\ell}$, there are as many sets (S) in the partition as there are series $\{k_i\}_{i=1}^{\ell}$ such that $\sum_{i=1}^{\ell} k_i = n$.
This means:

$$\left(\sum_{i=1}^{\ell} a_i\right)^n = \sum_{k_1+\cdots+k_\ell} \binom{n}{k_1,\ldots,k_\ell} \cdot a_1^{k_1} \cdots a_\ell^{k_\ell}$$

As required.

■

Theorem 1.2.8 (The Binomial Theorem):

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

Proof:

By the **The Multinomial Theorem**, this is equal to:

$$(a+b)^n = \sum_{k_1+k_2=n} \binom{n}{k_1, k_2} a^{k_1} \cdot b^{k_2}$$

But $k_1 + k_2 = n$ if and only if $k_2 = n - k_1$. That is, we can construct a simple bijection from the set of k_1, k_2s to the set $\{0, \ldots, n\}$ (by mapping (k_1, k_2) to k_1).
So instead we can sum over $\{0, \ldots, n\}$:

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

And by **proposition 1.2.5**, this is equal to:

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

As required.

■

2 Discrete Probability Spaces

2.1 Introduction to Probability Spaces

Definition 2.1.1:

A **probability space** is a triplet:

$$(\Omega, \mathcal{F}, \mathbb{P})$$

Where:

- Ω is a set called the **sample space**. Intuitively it is the set of all outcomes of a trial/experiment.
- $\mathcal{F}$ is a subset of $\mathcal{P}(\Omega)$, and its elements are called **events**. $\mathcal{F}$ must satisfy the following:
  - $\Omega \in \mathcal{F}$
  - If $\{A_i\}_{i=1}^{\infty} \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
  - If $A \in \mathcal{F}$, then $\Omega \setminus A \in \mathcal{F}$.

  Note:

  Note that these requirements also imply that:
  - $\varnothing \in \mathcal{F}$, since $\varnothing = \Omega \setminus \Omega$.
  - And if $\{A_i\}_{i=1}^{\infty} \in \mathcal{F}$ then:

    $$\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$$

    Since this is the complement of the union of the complements of A_i.

- $\mathbb{P}$ is the **probability function**, a function:

  $$\mathbb{P}\colon \mathcal{F} \longrightarrow [0, \infty)$$

  Which satisfies the following:
  - $\mathbb{P}(\Omega) = 1$
  - If $\{A_i\}_{i=1}^{\infty} \in \mathcal{F}$ are (pairwise) disjoint, then:

    $$\mathbb{P}\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

    Note that this is a **countably infinite** sum.

Proposition 2.1.2:

The following are true:

(1) $\mathbb{P}(\varnothing) = 0$

(2) If $\{A_i\}_{i=1}^{n} \in \mathcal{F}$ are disjoint, then:

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

This is different than the requirement on $\mathbb{P}$ since the sum is finite here.

(3) If A is a subset of B (and both are events), then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(4) If A is a subsetof B, then $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$

(5) If A is an event, then $\mathbb{P}(A) \in [0,1]$. This means that $\mathbb{P}$ can be thought of as a function to $[0,1]$ instead of as a function to $[0,\infty)$.

(6) $\mathbb{P}(A) = \mathbb{P}(A^c)$ (complements are relative to Ω).

Proof:

(1) We can define a series $\{A_i\}_{i=1}^{\infty}$ by $A_i = \varnothing$. Then they are all pairwise disjoint and their union is also $\varnothing$. So:

$$\mathbb{P}(\varnothing) = \mathbb{P}\left(\bigsqcup_{i=1}^{\infty} \varnothing\right) = \sum_{i=1}^{\infty} \mathbb{P}(\varnothing)$$

So we get:

$$\sum_{i=1}^{\infty} \mathbb{P}(\varnothing) = 0$$

Which means that $\mathbb{P}(\varnothing) = 0$ (as otherwise the sum doesn't converge).

(2) Let's define an infinite series $\{B_i\}_{i=1}^{\infty}$ like so:
For $i \leq n$, we define $B_i = A_i$. Otherwise, $B_i = \varnothing$.
Then $\{B_i\}_{i=1}^{\infty}$ is still pairwise disjoint and its union is $\bigsqcup_{i=1}^{n} A_i$, so:

$$\mathbb{P}\left(\bigsqcup_{i=1}^{n} A_i\right) = \mathbb{P}\left(\bigsqcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \sum_{i=1}^{n} \mathbb{P}(A_i) + \sum_{i=n+1}^{\infty} \mathbb{P}(\varnothing) = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

As required.

(3) We know that $B = A \sqcup (B \setminus A)$, and so:

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$$

And since $\mathbb{P}(B \setminus A) \geq 0$, we know $\mathbb{P}(A) \leq \mathbb{P}(B)$, as required.

(4) Similar to above, we see:

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$$

Which means that:

$$\mathbb{P}(B) - \mathbb{P}(A) = \mathbb{P}(B \setminus A)$$

As required.

(5) Since we know $\varnothing \subseteq A \subseteq \Omega$, this means:

$$\mathbb{P}(\varnothing) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) \implies 0 \leq \mathbb{P}(A) \leq 1$$

As required.

(6) We know that $\Omega = A \sqcup A^c$, so:

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

Which means that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, as required.

■

2.2 The Basics of Probability Spaces

Note:

Given the sum:

$$\sum_{x\in X} f(x)$$

Where X is potentially uncountable, we define it to be:

$$\sum_{x\in X} f(x) := \sum_{x\in \text{supp}_f(X)} f(x)$$

Where $\text{supp}_f(X)$ is the **support** of f on X, defined to be:

$$\text{supp}_f(X) := \{x \in X \mid f(x) \neq 0\}$$

Note that the sum must still be absolutely convergant, as we're summing in any order.

Definition 2.2.1:

A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is **discrete** if there exists a function:

$$P\colon \Omega \longrightarrow [0,1]$$

Such that for every event A:

$$\mathbb{P}(A) = \sum_{\omega\in A} P(\omega)$$

Definition 2.2.2:

A (finite) discrete probability space is **uniform** if for every $\omega \in \Omega$:

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

Proposition 2.2.3:

If $(\Omega, \mathcal{F}, \mathbb{P})$ is a uniform probability space, then:

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

Proof:

This is quite simple to prove. Notice:

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

As required. ■

Example:

Suppose you live on the planet Ekato-Karpouzia, which has k days in its year.
You're in a class with n people, what is the probability that at least two of the people in the class have the same birthday? (Assuming $n \leq k$.)
This is called the **birthday problem** or the **birthday paradox**, since the probability of this is higher than one would expect.

Here, we can define Ω to be the set of all functions from $[n]$ to $[k]$, where $f(i)$ gives the birthday of the ith person. We also know that the probability space must be uniform, since the probability of having one distribution of birthdays is the same as having another (this isn't the case in real life, but this is a good enough assumption for us). Notice that the event that there are at least two people in the class with the same birthday is the complement of the event that no one in the class shares a birthday with anyone else. This event is the set of all injective functions from $[n]$ to $[k]$. We know there are $\frac{k!}{(k-n)!}$ of those, which means the probability that everyone has a different birthday is:

$$\frac{k!}{(k-n)! \cdot k^n}$$

This is the probability of the complement of our goal event, which means the probability we're looking for is:

$$1 - \frac{k!}{(k-n)! \cdot k^n}$$

Now suppose that E. Karpouzia has the same number of days in a year as us, 365. That is, $k = 365$. If this were the case, even with 50 people, there'd be a probability greater than 0.97.

Theorem 2.2.4 (Law of Total Probability Version One):

If $\{A_i\}_{i \in I} \in \mathcal{F}$ is a countable partition of Ω, that is:

$$\bigsqcup_{i \in I} A_i = \Omega$$

Then for every event B:

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B \cap A_i)$$

Proof:

Notice that:

$$\bigsqcup_{i \in I} B \cap A_i = B \cap \left(\bigsqcup_{i \in I} A_i \right) = B \cap \Omega = B$$

Which means that:

$$\mathbb{P}(B) = \mathbb{P}\left(\bigsqcup_{i \in I} B \cap A_i \right)$$

And since I is countable, this is equal to:

$$\implies \mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B \cap A_i)$$

As required.

■

Example:

An n-sided die and a k-sided die are rolled. What is the probability the result of the n-sided die is less than that of

the k-sided die?

In this case, $\Omega = [n] \times [k]$, where $(x, y) \in \Omega$ corresponds to a result of x on the n-sided die and y on the k-sided die. The event we're trying to compute a probability for is:

$$B = \{(x, y) \in \Omega \mid x < y\}$$

Let's define for $a \in [k]$:

$$A_a = [n] \times \{a\} = \{(x, a) \in \Omega\}$$

Thus:

$$\mathbb{P}(B) = \sum_{a=1}^{k} \mathbb{P}(B \cap A_a)$$

$B \cap A_a = \{(x, a) \mid x < a\}$, which means that $|B \cap A_a| = a - 1$ (as there are $a - 1$ choices for x). Since the probability is uniform (the probability of rolling any two numbers is equal), this means:

$$\mathbb{P}(B \cap A_a) = \frac{|B \cap A_a|}{|\Omega|} = \frac{a-1}{n \cdot k}$$

Therefore:

$$\mathbb{P}(B) = \sum_{a=1}^{k} \frac{a-1}{n \cdot k} = \frac{1}{nk} \cdot \sum_{a=0}^{k-1} a = \frac{1}{nk} \cdot \frac{k}{2} \cdot (k-1) = \frac{k-1}{2n}$$

Lemma 2.2.5:

If $\{A_i\}_{i=1}^n$ are events, then:

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

Proof:

We can prove this through induction on $n = 1$. The base case for $n = 1$ is trivial. We will also prove it for $n = 2$.
Base case $n = 2$:
Notice that:

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1 \sqcup (A_2 \setminus A_1)) = \mathbb{P}(A_1) + \mathbb{P}(A_2 \setminus A_1)$$

And since $A_2 \setminus A_1$ is a subset of A_2, its probability is less than the probability of A_2. So we get:

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2 \setminus A_1) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2)$$

As required.
Inductive step:
Suppose the hypothesis is true for n. Let $\{A_i\}_{i=1}^{n+1}$ be events. Then:

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

And as we proved for $n = 2$, this is less than:

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

And by our inductive hypothesis, $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$, so this is less than:

$$\leq \sum_{i=1}^{n} \mathbb{P}(A_i) + \mathbb{P}(A_{n+1}) = \sum_{i=1}^{n+1} \mathbb{P}(A_i)$$

As required.

■

Definition 2.2.6:

A series of sets $\{A_i\}_{i=1}^{\infty}$ is **increasing** if for every $n \in \mathbb{N}_1$, A_n is a subset of A_{n+1}. And it is **decreasing** if for every $n \in \mathbb{N}_1$, A_n is a superset of A_{n+1}.

Theorem 2.2.7:

If $\{A_i\}_{i=1}^{\infty}$ is increasing, then:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

Proof:

Let's define $B_n := A_n \setminus A_{n-1}$.

Claim:

$\{B_n\}_{n=1}^{\infty}$ is pairwise disjoint.
Suppose $\omega \in B_n$, then $\omega \notin A_{n-1}$. Since $\{A_i\}_{i=1}^{\infty}$ is increasing, this means that for every $k < n$, we know $\omega \notin A_k$. And since $B_k \subseteq A_k$, this means that for every $k < n$, $\omega \notin B_k$.
Suppose that for some $k > n$, $\omega \in B_k$. But this would mean $\omega \notin B_n$ (by above), which is a contradiction.
So for every $k \neq n : \omega \notin B_k$. So $\{B_n\}_{n=1}^{\infty}$ is disjoint.

Claim:

$$\bigsqcup_{k=1}^{n} B_k = A_n$$

Suppose $\omega \in \bigsqcup B_k$, then there exists some $k \leq n$ such that $\omega \in B_k$. And since $B_k \subseteq A_k$, this means $\omega \in A_k$.
Now suppose ωA_n. There must be some k such that $\omega \in A_k$ and $\omega \notin A_{k-1}$ (take $k = \min\{m \leq n \mid \omega \in A_m\}$). This means $\omega \in A_k \setminus A_{k-1} = B_k$, and $k \leq n$, so $\omega \in \bigsqcup B_k$.
So an element is in one union if and only if it is in A_n, and therefore they are equal.

It follows that:

$$\bigsqcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$$

Since:

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} \bigsqcup_{k=1}^{n} B_k = \bigsqcup_{n=1}^{\infty} B_k$$

So:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigsqcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n)$$

Now recall that by definition:

$$\sum_{k=1}^{\infty} \mathbb{P}(B_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(B_k) = \lim_{n\to\infty} \mathbb{P}\left(\bigsqcup_{k=1}^{n} B_k\right)$$

And by our previous claim, this is equal to:

$$= \lim_{n\to\infty} \mathbb{P}(A_n)$$

As required.

■

Corollary 2.2.8:

If $\{A_i\}_{i=1}^{\infty}$ are events which are decreasing, then:

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

Proof:

Notice that since $A_n \supseteq A_{n+1}$, we know that $A_n^c \subseteq A_{n+1}^c$, so $\left\{A_i^c\right\}_{i=1}^{\infty}$ is an increasing series. Furthermore:

$$\bigcap_{n=1}^{\infty} A_n = \left(\bigcup_{n=1}^{\infty} A_n^c\right)^c$$

So:

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = 1 - \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n^c\right)$$

And since $\left\{A_i^c\right\}_{i=1}^{\infty}$ is increasing, by the above theorem:

$$= 1 - \lim_{n\to\infty} \mathbb{P}(A_n^c) = \lim_{n\to\infty} 1 - \mathbb{P}(A_n^c) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

As required.

■

Note:

These two theorems will become very important in the future when we discuss general probability spaces. So keep them in mind.

Theorem 2.2.9:

If I is countable then:

$$\mathbb{P}\left(\bigcup_{i\in I} A_i\right) \leq \sum_{i\in I} \mathbb{P}(A_i)$$

Proof:

We already proved the finite case in **lemma 2.2.5**, so all that remains is to prove the countably finite case. That is, we need to prove:

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

Let:

$$B_n := \bigcup_{k=1}^{n} A_k$$

Then $\{B_n\}_{n=1}^{\infty}$ is increasing, and:

$$\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$$

So by the previous theorem:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right)$$

And by **lemma 2.2.5**:

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} \mathbb{P}(A_k)$$

So:

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) \leq \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(A_k) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Which means that all in all:

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

As required.

■

2.3 The Inclusion-Exclusion Principle

Theorem 2.3.1 (The Inclusion-Exclusion Principle):

If $\{A_i\}_{i=1}^n$ are events, for every $I \leq [n]$, let:

$$A_I := \bigcap_{i \in I} A_i$$

Then:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \cdot \mathbb{P}(A_I)$$

Proof:

We will prove this through induction on n.
First base case: If $n = 1$, this is trivial.
Second base case: If $n = 2$, then notice that:

$$A \cup B = (A \setminus B) \sqcup B$$

And we know that $A \setminus B = A \setminus (A \cap B)$, and since $A \cap B$ is a subset of A:

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

So all in all we get:

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

As required.
Inductive step: Suppose this is true for n, then let $\{A_i\}_{i=1}^{n+1}$ be events. We know:

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

Which is equal to, by our second base case:

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

If we let $B_i := A_i \cap A_{n+1}$, notice that for every $\varnothing \neq I \subseteq [n]$:

$$B_I = \bigcap_{i \in I} A_i \cap A_{n+1} = A_{n+1} \cap \bigcap_{i \in I} A_i = A_{n+1} \cap A_I$$

So by our inductive hypothesis, this is equal to:

$$= \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \mathbb{P}(A_I) + \mathbb{P}(A_{n+1}) - \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \mathbb{P}(A_{n+1} \cap A_I)$$

Notice that:

$$\mathbb{P}(A_{n+1}) - \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \mathbb{P}(A_{n+1} \cap A_I) = \sum_{I \subseteq [n]} (-1)^{|I \cup \{n+1\}|+1} \mathbb{P}\left(A_{I \cup \{n+1\}}\right)$$

Since $(-1)^{|I \cup \{n+1\}|+1} = -1^{|I|+1}$ (so we're rewritting the minus before the sum), and when $I = \varnothing$, the element in the sum is just:

$$(-1)^{|\{n+1\}|+1} \cdot \mathbb{P}\left(A_{\{n+1\}}\right) = \mathbb{P}(A_{n+1})$$

So we entered A_{n+1} into the sum.
And this, in turn, is just equal to:

$$\sum_{\substack{I \subseteq [n+1] \\ n+1 \in I}} (-1)^{|I|+1} \cdot \mathbb{P}(A_I)$$

So, we get:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \cdot \mathbb{P}(A_I) + \sum_{\substack{I \subseteq [n+1] \\ n+1 \in I}} (-1)^{|I|+1} \cdot \mathbb{P}(A_I)$$

The first sum is summing over $\varnothing \neq I \subseteq [n+1]$ where $n+1 \notin I$, and the second is summing over $\varnothing \neq I \subseteq [n+1]$, where $n+1 \in I$. So they form a reordering of the sum of $\varnothing \neq I \subseteq [n+1]$, so their sum is equal to:

$$= \sum_{\varnothing \neq I \subseteq [n+1]} (-1)^{|I|+1} \cdot \mathbb{P}(A_I)$$

As required.

■

Note:

This is called the **Inclusion-Exclusion Principle** since we first add the probabilities of the events, then subtract the probabilities we double-counted (the intersections), then add back the probabilities which we double-counted in the previous step, and so on.
So this is a process of including, and then excluding.

Corollary 2.3.2:

If $\{A_i\}_{i=1}^{n}$ are finite sets, then:

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \cdot |A_I|$$

Proof:

Let $\Omega := \bigcup_{i=1}^{n} A_i$, and we define:

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

Thus if we define for every $A \subseteq \Omega$:

$$\mathbb{P}(A) = \sum_{\omega \in A} P(\omega)$$

$(\Omega, \mathcal{P}\Omega, \mathbb{P})$ forms a uniform probability space, so:

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

This means that:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \frac{\left|\bigcup_{i=1}^{n} A_i\right|}{|\Omega|}$$

But we know by the **The Inclusion-Exclusion Principle** that:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \cdot \mathbb{P}(A_I) = \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \cdot \frac{|A_I|}{|\Omega|}$$

Which means that:

$$\frac{\left|\bigcup_{i=1}^{n} A_i\right|}{|\Omega|} = \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \cdot \frac{|A_I|}{|\Omega|}$$

And multiplying both sides by $|\Omega|$ gives:

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \cdot |A_I|$$

As required.

■

Using the inclusion-exclusion principle, we can prove some pretty nifty stuff.

Example:

How many permutations of $[n]$ are there which have no stable points? (A point i is a stable point if $\sigma(i) = i$.)

We can start by asking the inverse of this question:

"How many permutations of S_n are there with at least 1 stable point?"

Which we can answer using inclusion-exclusion.
Let A be the set of all permutations in S_n with no stable points. That is:

$$A \coloneqq \{\sigma \in S_n \mid \forall i \in [n\} : \sigma(i) \neq i]$$

So we're going to try and find the cardinality of A^c instead:

$$A^c = \{\sigma \in S_n \mid \exists i \in [n\} : \sigma(i) = i]$$

We can then define the set B_i to be the set of all permutations where i is a stable point:

$$\forall i \in [n] : B_i \coloneqq \{\sigma \in S_n \mid \sigma(i) = i\}$$

This means that:

$$A^c = \bigcup_{i=1}^{n} B_i$$

Notice that for an indexing set $I \subseteq [n]$ the cardinality of B_I, where B_I is defined as:

$$B_I \coloneqq \bigcap_{i \in I} B_i$$

Is equal to $(n - |I|)!$. This is because we can create a simple bijection from $S_{[n]\setminus I}$ to B_I. Finding this bijection is simple and left as an exercise to the reader.
So, by inclusion-exclusion:

$$|A^c| = \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \cdot |B_I|$$

We can partition this into sums summing over cardinalities of I:

$$= \sum_{k=1}^{n} \sum_{\substack{I \subseteq [n] \\ |I|=k}} (-1)^{k+1} \cdot |B_I| = \sum_{k=1}^{n} \sum_{\substack{I \subseteq [n] \\ |I|=k}} (-1)^{k+1} \cdot (n-k)!$$

And since there are $\binom{n}{k}$ subsets of $[n]$ of cardinality k, this is equal to:

$$= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \cdot (n-k)!$$

And notice that $\binom{n}{k} \cdot (n-k)! = \frac{n!}{k!}$, so this is equal to:

$$= n! \cdot \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!}$$

Since $|A| = |S_n| - |A^c| = n! - |A^c|$, this means that:

$$|A| = n! - n! \cdot \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} = n! \cdot \left(1 - \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!}\right)$$

Notice that:

$$1 - \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} = 1 + \sum_{k=1}^{n} \frac{(-1)^{k}}{k!} = \sum_{k=0}^{n} \frac{(-1)^{k}}{k!}$$

So we have that:

$$|A| = n! \cdot \sum_{k=0}^{n} \frac{(-1)^{k}}{k!}$$

Note:

Notice that:

$$\frac{|A|}{n!} = \sum_{k=0}^{n} \frac{(-1)^{k}}{k!}$$

This is equal to the probability of uniformly choosing a permutation with no stable points from S_n.
This probability has the interesting property that:

$$\lim_{n\to\infty} \frac{|A|}{n!} = \sum_{k=0}^{\infty} \frac{(-1)^{k}}{k!}$$

Which is the Taylor Series of e^x at $x = -1$! So:

$$\lim_{n\to\infty} \frac{|A|}{n!} = e^{-1} = \frac{1}{e}$$

Which means that as you increase n, the probability of choosing a permutation with no stable points approaches $\frac{1}{e}$.
This won't be the last time you see e in probability...

Example:

Let's continue on the same thought as the previous example, but generalizing a bit. How many permutations with exactly k stable points are there?

We could solve this in a similar fashion to the previous example, but what would doing the same thing again really achieve?
Instead, let's use the previous example to solve this one.
Let A be the set of all permutations with exactly k stable points, and let for all $I \subseteq [n]$, B_I be the set of all permutations where every $i \in I$ is a stable point and every $i \notin I$ is not:

$$B_I \coloneqq \{\sigma \in S_n \mid \forall i \in I : \sigma(i) = i \ \wedge \ \forall i \notin I : \sigma(i) \neq i\}$$

Notice that for every $I \neq J \subseteq [n]$, B_I and B_J are disjoint. Furthemore, notice that:

$$A = \bigsqcup_{\substack{I \subseteq [n] \\ |I| = k}} B_I$$

So now the question boils down to finding the cardinality of B_I.
Let C be the set of all permutations of $[n-k]$ which have no stable points. By our previous example, we know that

$$|C| = (n-k)! \cdot \sum_{i=0}^{n-k} \frac{(-1)^{i}}{i!}$$

Furthermore, we can construct a bijection from B_I to C:

$$f : B_I \longrightarrow C$$

Since $|I| = k$, there exists a bijection g_I from $[n-k]$ to I^c. So we can define:

$$f(\sigma) = \tau$$

Where for every $m \in [n-k]$:

$$\tau(m) = g_I^{-1}(\sigma(g_I(m)))$$

(This just means $f(\sigma) = g_I^{-1} \circ \sigma \circ g_I$ if we ignore domains and codomains.)
This is well defined since $g_I(m) \in I^c$, and since $\sigma \in B_I$, every element of I^c is not a stable point, so:

$$\sigma(g_I(m)) \neq g_I(m)$$

And since g_I^{-1} is bijective, and therefore injective, this means:

$$\tau(m) = g_I^{-1}(\sigma(g_I(m))) \neq g_I^{-1}(g_I(m)) = m$$

So for every $m \in [n-k]$:

$$\tau(m) \neq m$$

Which means that τ has no stable points, so $m \in C$.
Also note that since $g_I(m) \notin I$, this means that $\sigma(g_I(m)) \notin I$ as well (since if it were in I, since σ is a permutation, it'd have to be a stable point), so we can take the g_I^{-1} of that.
This is also obviously injective since if:

$$f(\sigma) = f(\pi) \implies g_I^{-1} \circ \sigma \circ g_I = g_I^{-1} \circ \pi \circ g_I \implies \sigma = \pi$$

And if $\tau \in C$, then for every $i \notin I$, we can define $\sigma(i) = g_I \circ \tau \circ g_I^{-1}(i)$, and if $i \in I$, then $\sigma(i) = i$. This is in B_I since every $i \in I$ is a stable point and:

$$g_I \circ \tau \circ g_I^{-1}(i) = i \iff \tau \circ g_I^{-1}(i) = g_I^{-1}(i)$$

Which is false since $\tau \in C$, so it has no stable points.
Thus

$$|B_I| = |C|$$

And so:

$$|A| = \sum_{\substack{I \subseteq [n] \\ |I|=k}} |B_I| = \sum_{\substack{I \subseteq [n] \\ |I|=k}} |C| = \binom{n}{k} \cdot |C| = \binom{n}{k} \cdot (n-k)! \cdot \sum_{i=0}^{n-k} \frac{(-1)^i}{i!} = \frac{n!}{k!} \cdot \sum_{i=0}^{n-k} \frac{(-1)^i}{i!}$$

Note:

Similarly, here if you look at:

$$\frac{|A|}{n!} = \frac{1}{k!} \cdot \sum_{i=0}^{n-k} \frac{(-1)^i}{i!}$$

This is the probability of choosing a permutation with exactly k stable points from S_n.
And its limit as n approaches infinity is:

$$\lim_{n \to \infty} \frac{|A|}{n!} = \frac{1}{k!} \cdot \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} = \frac{1}{k!} \cdot e^{-1} = \frac{1}{k! \cdot e}$$

Yet another example of the usage of the inclusion-exclusion principle arises in number theory.

Definition:

The **Euler Totient Function** is a function:

$$\varphi : \mathbb{N}_1 \longrightarrow \mathbb{N}_1$$

Where $\varphi(n)$ is equal to the number of numbers in $[n]$ which are coprime with n.
(Two numbers are coprime if they share no common divisior: in other words, if their greatest common divisor is 1.)

The Euler Totient Function has many uses in number theory, and has some very interesting properties which won't be covered here.

Lemma:

$$\prod_{k=1}^{n}(a_k + b_k) = \sum_{I \subseteq [n]} \prod_{i \in I} a_i \cdot \prod_{i \notin I} b_i$$

Proof:

This can be shown simply using some combinatorics. By writing out the product, see that:

$$\prod_{k=1}^{n}(a_k + b_k) = (a_1 + b_1) \cdot (a_2 + b_2) \cdots (a_n + b_n)$$

We can expand this product by choosing from each parentheses a_k or b_k and multiplying all these choices together. Let I be the set of indexes for the parentheses where we chose a_i (so for $a_1 \cdot a_2 \cdot a_n$, $I = \{1, 2, n\}$) so I^c is the set of indexes where we chose b_i.
For every I like this, the coefficient we are adding to the sum is:

$$\prod_{i \in I} a_i \cdot \prod_{i \notin I} b_i$$

Any subset $I \subseteq [n]$ is a valid choice, so we can sum over $I \subseteq [n]$:

$$\prod_{k=1}^{n}(a_k + b_k) = \sum_{I \subseteq [n]} \prod_{i \in I} a_i \cdot \prod_{i \notin I} b_i$$

As required.

Note:

This isn't really a rigorous proof, a rigorous proof can be done simply with induction. But the purpose of this is not to talk about induction, rather combinatorics and probability. Therefore, the "rigorous" proof of this lemma is left as an exercise to the reader.

Example:

Suppose $n = p_1^{k_1} \cdots p_t^{k_t}$ where p_i is prime. Then:

$$\varphi(n) = n \cdot \prod_{i=1}^{t}\left(1 - \frac{1}{p_i}\right)$$

In other words:

$$\varphi(n) = n \cdot \prod_{p \mid n}\left(1 - \frac{1}{p}\right)$$

I highly urge you to try and prove this yourself, it provides a good example of how studying probability can be useful in other fields, and not just with things related to probability.

Let's define:

$$A \coloneqq \{m \in [n] \mid \gcd(n, m) = 1\}$$

So $\varphi(n) = |A|$. ($\gcd(n, m)$ is the greatest common divisor of n and m, so $\gcd(n, m) = 1$ means that n and m are coprime.)
Now let's look at A's complement: the set of all numbers which share a divisior with n. This is true if and only if they are divisible by some q_i. So:

$$A^c = \{m \in [n] \mid \exists i : q_i \mid m\}$$

Now, we can define forall $1 \leq i \leq t$:

$$B_i \coloneqq \{m \in [n] \mid q_i \mid n\}$$

So:

$$A^c = \bigcup_{i=1}^{t} B_{q_i}$$

Which means that by the inclusion-exclusion principle:

$$|A^c| = \sum_{\varnothing \neq I \subseteq [n]} (-1)^{|I|+1} \cdot |B_I|$$

So what is the cardinality of B_I? Well, we know that:

$$B_I = \{m \in [n] \mid \forall i \in I : q_i \mid m\}$$

And since q_i are all prime, every q_i divides m if and only if the product of q_i over I divides m, so:

$$B_I = \left\{ m \in [n] \;\middle|\; \left(\prod_{i \in I} q_i \right) \mid m \right\}$$

So B_I is all the multiples of $\prod_{i \in I} q_i$ in $[n]$:

$$B_I = \left\{ k \cdot \prod_{i \in I} q_i \in [n] \;\middle|\; k \in \mathbb{N}_1 \right\}$$

And recall that $\prod_{i \in I} q_i$ divides n, since the prime factorization of n includes every q_i. This means that:

$$|B_I| = \frac{n}{\prod_{i \in I} q_i}$$

(Since there are these many choices for k. Alternatively, just build a bijection from $\left[\frac{n}{\prod q_i}\right]$ to B_I. This bijection is simple and is again, left as an exercise to the reader.)
So we have that:

$$|A^c| = n \cdot \sum_{\varnothing \neq I \subseteq [n]} \frac{(-1)^{|I|+1}}{\prod_{i \in I} q_i}$$

And so:

$$|A| = n - n \cdot \sum_{\varnothing \neq I \subseteq [n]} \frac{(-1)^{|I|+1}}{\prod_{i \in I} q_i} = n \left(1 - \sum_{\varnothing \neq I \subseteq [n]} \frac{(-1)^{|I|+1}}{\prod_{i \in I} q_i} \right) = n \cdot \sum_{I \subseteq [n]} \frac{(-1)^{|I|}}{\prod_{i \in I} q_i}$$

(Since the empty product is equal to 1.)
Now, notice that this is equal to:

$$n \cdot \sum_{I \subseteq [n]} \prod_{i \in I} \left(-\frac{1}{q_i} \right) = n \cdot \sum_{I \subseteq [n]} \prod_{i \in I} \left(-\frac{1}{q_i} \right) \cdot \prod_{i \notin I} 1$$

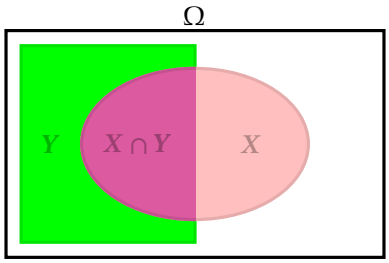
Which, by our lemma above, is equal to:

$$n \cdot \prod_{i=1}^{t} \left(1 - \frac{1}{q_i} \right)$$

As required.

2.4 Conditional Probability and Independence

A lot of probability questions are in the form of "*What is the probability of X if Y?*" So we want to try and formulate this question mathematically.
The idea is that if we know Y (where Y is some event), then we can think of Y as the new sample space. Then the probability of X would be the probability of $X \cap Y$ divided by the probability of Y.
The rationale behind this may become more clear with the following illustration:



We want to focus our attention on $X \cap Y$ within Y.

Definition 2.4.1:

If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and B is an event in $\mathcal{F}$ such that $\mathbb{P}(B) > 0$, we define the **conditional probability function** of B to be a function:

$$\mathbb{P}(\cdot \mid B) : \mathcal{F} \longrightarrow [0,1]$$

Where $\mathbb{P}(A \mid B)$ is defined as:

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

Another notation for $\mathbb{P}(A \mid B)$ is $\mathbb{P}_B(A)$.

Proposition 2.4.2 (Baye's Law):

$$\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A) \cdot \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$$

Proof:

This is quite simple, notice:

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

As required. ■

Theorem 2.4.3 (Law of Total Probability Version Two):

If $\{A_i\}_{i \in I} \in \mathcal{F}$ is a countable partition of Ω, and for every $i \in I$, $\mathbb{P}(A_i) > 0$, then for every event $B \in \mathcal{F}$:

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)$$

Proof:

By the **Law of Total Probability Version One**, we know:

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B \cap A_i)$$

And since $\mathbb{P}(B \cap A_i) = \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)$, we see:

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)$$

As required.

■

Lemma 2.4.4:

If A, B, and C are events, such that $\mathbb{P}(B), \mathbb{P}(C) > 0$, then:

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

Proof:

We know:

$$\mathbb{P}(A \mid B \mid C) = \mathbb{P}_C(A \mid B) = \frac{\mathbb{P}_C(A \cap B)}{\mathbb{P}_C(B)} = \frac{\mathbb{P}(A \cap B \cap C)/\mathbb{P}(C)}{\mathbb{P}(B \cap C)/\mathbb{P}(C)} = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)}$$

And on the other hand, we know:

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

As required.

■

Theorem 2.4.5:

If $\{A_i\}_{i=1}^n \in \mathcal{F}$ are events, then:

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbb{P}\left(A_i \,\middle|\, \bigcap_{j=1}^{i-1} A_j\right)$$

This is the mathematical concept behind the commonly taught "tree method" for computing the probability of intersections.
When $i = 1$, we must define $\bigcap_{j=1}^{i-1} A_j = \Omega$, since conditional probability is not defined on $\varnothing$, as it has zero probability.

Proof:

We will prove this through induction on n.
Base case: $n = 1$.
Then the product is equal to simply:

$$\mathbb{P}(A_1 \mid \Omega) = \mathbb{P}(A_1)$$

As required.
Base case: $n = 2$
So we need to show that:

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \mid A_1)$$

And this is true by the definition of conditional probability.
Inductive step: We know:

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

Which is equal to, by our inductive hypothesis:

$$= \prod_{i=1}^{n-1} \mathbb{P}\left(A_i \,\middle|\, \bigcap_{j=1}^{i-1} A_j\right) \cdot \mathbb{P}\left(A_n \cap A_{n+1} \,\middle|\, \bigcap_{j=1}^{n-1} A_j\right)$$

And by our base case for $n = 2$, we know that:

$$\mathbb{P}\left(A_n \cap A_{n+1} \,\middle|\, \bigcap_{j=1}^{n-1} A_j\right) = \mathbb{P}\left(A_n \,\middle|\, \bigcap_{j=1}^{n-1} A_j\right) \cdot \mathbb{P}\left(A_{n+1} \,\middle|\, A_n \,\middle|\, \bigcap_{j=1}^{n-1} A_j\right)$$
$$= \mathbb{P}\left(A_n \,\middle|\, \bigcap_{j=1}^{n-1} A_j\right) \cdot \mathbb{P}\left(A_{n+1} \,\middle|\, \bigcap_{j=1}^{n} A_j\right)$$

So all in all:

$$\mathbb{P}\left(\bigcap_{i=1}^{n+1} A_i\right) = \prod_{i=1}^{n-1} \mathbb{P}\left(A_i \,\middle|\, \bigcap_{j=1}^{i-1} A_j\right) \cdot \mathbb{P}\left(A_n \,\middle|\, \bigcap_{j=1}^{n-1} A_j\right) \cdot \mathbb{P}\left(A_{n+1} \,\middle|\, \bigcap_{j=1}^{n} A_j\right)$$

And this is just equal to:

$$= \prod_{i=1}^{n+1} \mathbb{P}\left(A_i \,\middle|\, \bigcap_{j=1}^{i-1} A_j\right)$$

As required.

■

Definition 2.4.6:

Two events, A and B, are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.
This is denoted as $A \perp\!\!\!\perp B$.

It is worth noting that independence is symmetric: if A and B are independent, B and A are independent. This is because intersection and multiplication is commutative.

Proposition 2.4.7:

The following are equivalent:

(1) A and B are independent.

(2) $\mathbb{P}(A \mid B) = \mathbb{P}(A)$

(3) $\mathbb{P}(B \mid A) = \mathbb{P}(B)$

Proof:

(1 $\implies$ 2) We know that:

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

As required.

(2 $\implies$ 3) By **Baye's Law** we know:

$$\mathbb{P}(B \mid A) = \mathbb{P}(A \mid B) \cdot \frac{\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(A) \cdot \frac{\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B)$$

As required.

(3 $\implies$ 1) We know:

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

Which means A and B are independent, as required.

■

Proposition 2.4.8:

(1) If $\mathbb{P}(A)$ is equal to 0 or 1 if and only if for every $B \in \mathcal{F}$, $A \perp\!\!\!\perp B$.

(2) An event A is independent of every every event if and only if it is independent of itself.

(3) If A and B are disjoint events with non-zero probability, they are not independent.

(4) If A is a subset of B, $\mathbb{P}(A) \neq 0$, and $\mathbb{P}(B) \neq 1$, then A and B are not independent.

(5) If A and B are independent, so are A^c and B.

Proof:

(1) ($\implies$) Suppose $\mathbb{P}(A) = 0$. Since $A \cap B \subseteq A$, that means $\mathbb{P}(A \cap B) \leq \mathbb{P}(A) = 0$, so $\mathbb{P}(A \cap B) = 0$. And since $\mathbb{P}(A) \cdot \mathbb{P}(B) = 0 \cdot \mathbb{P}(B) = 0$, $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$, as required.
Now suppose $\mathbb{P}(A) = 1$. We know then that $\mathbb{P}(A^c) = 0$. We also know that:

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

By the union bound, $\mathbb{P}(A^c \cup B^c) \leq \mathbb{P}(A^c) + \mathbb{P}(B^c) = \mathbb{P}(B^c)$. But on the other hand, since $B^c \subseteq A^c \cup B^c$, $\mathbb{P}(B^c) \leq \mathbb{P}(A^c \cup B^c)$, so $\mathbb{P}(A^c \cup B^c) = \mathbb{P}(B^c)$.
Therefore:

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

As required.
($\impliedby$) Since A is independent of every event, it must be independent of itself, so:

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2$$

This means that $\mathbb{P}(A)$ is equal to 0 or 1, as required.

(2) As shown in the above proof, A is independent of itself if and only if $\mathbb{P}(A)$ is 0 or 1. And by above, this is equivalent to A being independent of every event.

(3) Since $A \cap B = \varnothing$, $\mathbb{P}(A \cap B) = 0$. But $\mathbb{P}(A), \mathbb{P}(B) > 0$, so $\mathbb{P}(A) \cdot \mathbb{P}(B) \neq 0$, so A and B are dependent.

(4) Since $A \subseteq B$, $\mathbb{P}(A \cap B) = \mathbb{P}(A)$. This equal to $\mathbb{P}(A) \cdot \mathbb{P}(B)$ if and only if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 1$, which they don't.

(5) We know

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

As required.

> **Note:**
> Using this, we can show that $A \perp\!\!\!\perp B$, $A^c \perp\!\!\!\perp B$, $A \perp\!\!\!\perp B^c$, and $A^c \perp\!\!\!\perp B^c$ are all equivalent.

Definition 2.4.9:

A set of events, $\{A_i\}_{i=1}^n$ is independent if for every set $I \subseteq [n]$:

$$\mathbb{P}(A_I) = \prod_{i \in I} \mathbb{P}(A_i)$$

(Recall that $A_I = \bigcap_{i \in I} A_i$.)
And for an arbitrary indexing set I, $\{A_i\}_{i \in I}$ is independent if for every finite subset $J \subset I$, $\left\{A_j\right\}_{j \in J}$ is independent.

Proposition 2.4.10:

Suppose $\{A_i\}_{i=1}^{\infty}$ is independent, then:

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

Proof:

Let:

$$B_m := \bigcap_{i=1}^{m} A_m$$

Which means that:

$$\bigcap_{m=1}^{\infty} B_m = \bigcap_{m=1}^{\infty} A_m$$

Furthermore, we know that B_m must be decreasing as $B_{m+1} = B_m \cap A_{m+1}$. Therefore, by **theorem 2.2.7**, we know that:

$$\lim_{m\to\infty} \mathbb{P}(B_m) = \mathbb{P}\left(\bigcap_{i=1}^{\infty} B_i\right)$$

And by the definition of B_m, this means

$$\lim_{m\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{m} A_i\right) = \mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right)$$

Since $\{A_i\}_{i=1}^{\infty}$ is independent:

$$\lim_{m\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{m} A_i\right) = \lim_{m\to\infty} \prod_{i=1}^{m} \mathbb{P}(A_i) = \prod_{i=1}^{\infty} \mathbb{P}(A_i)$$

So:

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

As required.

■

Definition 2.4.11:

Suppose $\{B_i\}_{i=1}^{n}$ is a set of events, and A is some other event. We say that A is independent of $\{B_i\}_{i=1}^{n}$ if for every set $I \subseteq [n]$, A and B_I are independent.

Proposition 2.4.12:

Suppose $\{A_i\}_{i=1}^{n}$ is a set of events. Then $\{A_i\}_{i=1}^{n}$ is independent if and only if every A_j is independent of $\{A_i\}_{i=1}^{n} \setminus \{A_j\}$.

Proof:

($\Longrightarrow$) Let $I \subseteq [n]$ such that $j \notin I$. We must show that A_j and A_I are independent. We know that:

$$\mathbb{P}\left(A_j \cap A_I\right) = \mathbb{P}\left(A_{I\cup\{j\}}\right) = \prod_{i\in I} \mathbb{P}(A_i) \cdot \mathbb{P}\left(A_j\right) = \mathbb{P}(A_I) \cdot \mathbb{P}\left(A_j\right)$$

As required.

($\Longleftarrow$) Suppose $I = \{i_1, \ldots, i_k\} \subseteq [n]$. Then:

$$\mathbb{P}(A_I) = \mathbb{P}\left(A_{i_1} \cap A_{I\setminus\{i_1\}}\right)$$

Since A_{i_1} is independent of $A_{I\setminus\{i_1\}}$, this is equal to:

$$= \mathbb{P}\left(A_{i_1}\right) \cdot \mathbb{P}\left(A_{I\setminus\{i_1\}}\right)$$

And then using induction on the size of I, we get that this is equal to:

$$= \mathbb{P}\left(A_{i_1}\right) \cdot \prod_{i \in I \setminus \{i_1\}} \mathbb{P}(A_i) = \prod_{i \in I} \mathbb{P}(A_i)$$

So $\{A_i\}_{i=1}^n$ is independent.

■

Definition 2.4.13:

A set $\{A_i\}_{i=1}^n$ is **pairwise independent** if for every relevant $i \neq j$, A_i and A_j are independent.

Note:

If a set is independent, it is also pairwise independent. This is since you can take the set $\{i, j\}$.
But the reverse is not true. This can be demonstrated with the following example:
Suppose we flip a coin k times, where k is odd. Our sample space Ω will be the set of all vectors $\boldsymbol{x} \in [0,1]^k$ which correspond to the result of the flips ($\boldsymbol{x}_i$ is the result of the ith flip, 1 is heads, etc.). Let's define the following events for all $i \in [k]$:

$$A_i = \{\boldsymbol{x} \in \Omega \mid \boldsymbol{x}_i = 1\}$$

And:

$$B = \left\{ \boldsymbol{x} \in \Omega \;\middle|\; \sum_{i=1}^{k} \boldsymbol{x}_i \equiv 0 \pmod 2 \right\}$$

(A_i is the event that the ith flip resulted in heads, B is the event that there is an even amount of heads.)
Notice that the probability of A_i and the probability of B are both $\frac{1}{2}$ (by symmetry), and:

$$A_i \cap B = \{\boldsymbol{x} \mid \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_{i-1} + 1 + \boldsymbol{x}_{i+1} \cdots + \boldsymbol{x}_k \equiv 0 \pmod 2\} = \left\{ \boldsymbol{x} \;\middle|\; \sum_{\substack{j=1 \\ j \neq i}} \boldsymbol{x}_j \equiv 1 \pmod 2 \right\} \cap \{\boldsymbol{x} \mid \boldsymbol{x}_i = 1\}$$

Both of these events are independent and have probability $\frac{1}{2}$, so the probability $\mathbb{P}(A_i \cap B) = \frac{1}{4}$. So A_i and B are independent. Furthermore, A_i and A_j are independent (this is trivial). So the set $\{A_1, \ldots, A_k, B\}$ is pairwise independent.
But:

$$A_{[k]} \cap B = \left\{ \boldsymbol{x} \;\middle|\; \forall i \in [k] : \boldsymbol{x}_i = 1, \sum_{i=1}^{k} \boldsymbol{x}_i \equiv 0 \pmod 2 \right\} = \{\boldsymbol{x} \mid \boldsymbol{x}_i = 1, k \equiv 0 \pmod 2\}$$

But since k is odd, $k \not\equiv 0 \pmod 2$, so the set is the empty set and therefore has probability 0. But A_i and B don't, so $\{A_1, \ldots, A_k, B\}$ is not independent.

2.5 Discrete Random Variables

Often times it will be tricky, unintuitive, and just downright unhelpful to work with a plain probability space. It can be hard to figure out exactly *how* you want to construct your probability space, and you'll probably end up using lots of annoying and incomprehensible notation in order to make a point. Fortunately, probability theory offers a sort of abstraction from the notion of a probability space. This abstraction comes in the form of something called a *random variable*.

Definition 2.5.1:

A **random variable** over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function:

$$X : \Omega \longrightarrow \mathbb{R}$$

Note:

There's actually a bit more nuance to this definition, but we'll return to it when we cover general probability spaces.

Of course the introduction of random variables as functions should be of no surprise. Mathematicians just *love* functions. It makes them wet.
But why is it called a random *variable*? After all it's hardly a variable, it's a function for god's sake! Well think of it this way: suppose we have a clinical test where we want to know the number of patients who survived. We can call this a variable X, and we can play around with it like it's a variable, but it's not a constant! It is actually a function, where given some result of the test ω, it gives us the number of survivors. This is a function. So while on the surface a random variable seems like a variable, and in fact it is helpful to think of them as variables in many cases, they are far from being constants (though they can be, just like how there are constant functions).

Definition 2.5.2:

The **probability distribution** of a random variable X is a function:

$$\mathbb{P}_X : \mathcal{P}(\mathbb{R}) \longrightarrow [0,1]$$

Defined like so:

$$\mathbb{P}_X(S) = \mathbb{P}(\{\omega \in \Omega\} \mid X(\omega) \in S)$$

This is the probability that X is in the set S.
Many times there are special cases of this, like $\mathbb{P}(X \leq a)$ is the probability that X is less than 5, etc. Of course, this is all notational.

Proposition 2.5.3:

For every random variable X over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $(\mathbb{R}, \mathcal{P}(\mathbb{R}), \mathbb{P}_X)$ is a probability space.

Proof:

We need to show firstly that $\mathbb{P}_X(\mathbb{R}) = 1$. This is trivial since:

$$\mathbb{P}_X(\mathbb{R}) = \mathbb{P}(\{\omega \in \Omega\} \mid X(\omega) \in \mathbb{R})$$

But every ω satisfies that $X(\omega) \in \mathbb{R}$, so this is equal to $\mathbb{P}(\Omega) = 1$.
Now suppose $\{A_i\}_{i=1}^{\infty} \in \mathcal{P}(\mathbb{R})$ are disjoint. Then:

$$\mathbb{P}_X\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\{\omega \in \Omega\} \,\middle|\, X(\omega) \in \bigsqcup_{i=1}^{\infty} A_i\right)$$

But notice that:

$$\left\{\omega \in \Omega \,\middle|\, X(\omega) \in \bigsqcup_{i=1}^{\infty} A_i\right\} = \bigsqcup_{i=1}^{\infty} \{\omega \in \Omega \mid X(\omega) \in A_i\}$$

So this is equal to:

$$= \mathbb{P}\left(\bigsqcup_{i=1}^{\infty} \{\omega \in \Omega \mid X(\omega) \in A_i\}\right) = \sum_{i=1}^{\infty} \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A_i\}) = \sum_{i=1}^{\infty} \mathbb{P}_X(A_i)$$

As required.

■

Definition 2.5.4:

A random variable X is **discrete** if its probability distribution is discrete. That means that there exists a function

$$P_X : \mathbb{R} \longrightarrow [0,1]$$

such that for every $A \subseteq \mathbb{R}$:

$$\mathbb{P}_X(A) = \sum_{x \in A} P_X(x)$$

We also denote:

$$\mathbb{P}(X = a) = P_X(a)$$

Definition 2.5.5:

Given an event A, we define the **indicator function** of A to be a random variable defined like so:

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

Definition 2.5.6:

A random variable X has a **bernoulli distribution** of $p \in [0,1]$ if:

$$\mathbb{P}_X(\{1\}) = p \quad \mathbb{P}_X(\{0\}) = 1 - p$$

This is indicated $X \sim \text{Ber}(p)$.
Random variables with a bernoulli distribution are also called **indicators**, as they *indicate* if an event occurred.

Note:

It should be obvious that indicators are discrete. This is since we can define $P_X(x)$ to be equal to $\mathbb{P}_X(\{x\})$.
But note that this isn't always true. Say we want to somehow take a random number from $[0,1]$, the probability that we take any arbitrary number is 0 since there is an uncountable number of numbers to choose from, so there can't be a discrete probability function. We currently don't have the tools to deal with this, so we'll return to it later.

Proposition 2.5.7:

$$\mathbb{1}_A \sim \text{Ber}(\mathbb{P}(A))$$

Proof:

We know that $\mathbb{P}_{\mathbb{1}_A}(\{1\}) = \mathbb{P}(\{\mathbb{1}_A = 1\}) = \mathbb{P}(A)$, as required. And $\mathbb{P}_{\mathbb{1}_A}(\{0\}) = \mathbb{P}(\{\mathbb{1} = 0\}) = \mathbb{P}(\{\omega \notin A\}) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, as required.

■.

Definition 2.5.8:

A discrete random variable X has a **uniform distribution** over a set S, denoted $X \sim \text{Unif}[S]$ if the probability of X being equal to any $x \in S$ is equal, and $\mathbb{P}(X \notin S) = 0$. That is, for every $x, y \in S : \mathbb{P}(X = x) = \mathbb{P}(X = y)$.

Since $\mathbb{P}(X \in S) = 1$, this means that $\sum_{x \in S} \mathbb{P}(X = x) = |S| \cdot \mathbb{P}(X = x) = 1$, so for every $x \in S$, we have that $\mathbb{P}(X = x) = \frac{1}{|S|}$.
Since we have defined a new mathematical object it is handy to define equivalence on it. For random variables we define two types of equivalence:

Definition 2.5.9:

Two random variables, X and Y, *over the same probability space* are **almost surely equal** if the probability that they equal one another is 1. This is denoted by $X \overset{as}{=} Y$. So:

$$X \overset{as}{=} Y \iff \mathbb{P}(X = Y) = 1 \iff \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = Y(\omega)\}) = 1$$

We say that two discrete random variables X and Y are **distributively equal** if they have the same distribution. This is denoted by $X \overset{d}{=} Y$. So $X \overset{d}{=} Y$ if and only if $\mathbb{P}_X = \mathbb{P}_Y$. Note that X and Y don't need to necessarily be defined over the same probability space.

It is trivial to see that if $X = Y$ (that is they are the same function), then $X \overset{as}{=} Y$ and $X \overset{d}{=} Y$. But is almost surely equivalence more powerful than distributive equivalence? The answer is yes.

Proposition 2.5.10:

If X is almost surely equal to Y, than X is distributively equal to Y.

Proof:

Let $S \subseteq \mathbb{R}$, we will show that $\mathbb{P}_X(S) = \mathbb{P}_Y(S)$. Notice that $\mathbb{P}(X \in S) = \mathbb{P}(X \in S, Y \in S) + \mathbb{P}(X \in S, Y \notin S)$. But

$$\{X \in S, Y \notin S\} = \{\omega \in \Omega \mid X(\omega) \in S, Y(\omega) \notin S\}$$

which is a subset of the set $\{\omega \in \Omega \mid X(\omega) \neq Y(\omega)\}$ which has a probability of 0 since X and Y are almost surely equal. So we have that $\mathbb{P}_X(S) = \mathbb{P}(X \in S, Y \in S)$, and by symmetry $\mathbb{P}_Y(S) = \mathbb{P}(X \in S, Y \in S)$, so we have that for every $S \subseteq \mathbb{R}$: $\mathbb{P}_X(S) = \mathbb{P}_Y(S)$.

qed

Proposition 2.5.11:

Suppose X and Y are two random variables and $f : \mathbb{R} \longrightarrow \mathbb{R}$ is any real function.

(1) If $X \overset{as}{=} Y$ then $f(X) \overset{as}{=} f(Y)$.

(2) If $X \overset{d}{=} Y$ then $f(X) \overset{d}{=} f(Y)$.

Note:

It would be more correct to say that $f \circ X = f \circ Y$ since we're considering the composition of functions, but in probability it is much more common to treat random variables in a way similar to numbers. This is a much more comfortable way to treat them and there's no issue with it.

Proof:

(1) We need to show that $\mathbb{P}(f(X) = f(Y)) = 1$. That is, $\mathbb{P}(\{\omega \in \Omega \mid f(X(\omega)) = f(Y(\omega))\}) = 1$. But this set is a superset of the set $\{\omega \in \Omega \mid X(\omega) = Y(\omega)\}$, which has probability 1, and thus so does this. So $f(X) \overset{as}{=} f(Y)$.

(2) We need to show that $\mathbb{P}_{f(X)} = \mathbb{P}_{f(Y)}$. Let $s \subseteq \mathbb{R}$, then:

$$\mathbb{P}_{f(X)}(S) = \mathbb{P}(f(X) \in S) = \mathbb{P}\left(X \in f^{-1}(S)\right) = \mathbb{P}\left(Y \in f^{-1}(S)\right) = \mathbb{P}(f(Y) \in S)$$

As required.

■

Notice that if we have two random variables X and Y, even if we know their distribution, we still can't know the distribution of random variables in the form of $f(X,Y)$, for example $X+Y$. This is demonstrated in the following example:

Example:

Suppose $X, Y \sim \text{Ber}\left(\frac{1}{2}\right)$. In one case, suppose X and Y are the results of two independent coin flips, then $X+Y$ has a distribution of:

$$\mathbb{P}(X+Y=x) = \begin{cases} \frac{1}{4} & x = 0,2 \\ \frac{1}{2} & x = 1 \end{cases}$$

since if $x = 0$ or 2 then both coins must be either heads or tails (one out of four possibilities), and if $x = 1$ then one coin must be heads and the other tails (two out of four possibilities).
But if $X = Y$ then $X + Y = 2X$ which has a distribution $\mathbb{P}(X=0) = \frac{1}{2}$ and $\mathbb{P}(X=2) = \frac{1}{2}$.

So we need an extra piece of information in order to discern more about how random variables interact with one another.

Definition 2.5.12:

A **joint probability vector** is a vector of random variables. So for example, we may have

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$$

And the **joint probability distribution** of a joint probability vector $\boldsymbol{X}$ is a function:

$$\mathbb{P}_{\boldsymbol{X}} : \mathcal{P}(\mathbb{R}^n) \longrightarrow [0,1]$$

Where for every $S \in \mathcal{P}(\mathbb{R}^n)$:

$$\mathbb{P}_{\boldsymbol{X}}(S) = \mathbb{P}((X_1, \ldots, X_n) \in S)$$

A random vector is discrete if there exists a function $P_{\boldsymbol{X}} : \mathbb{R}^n \longrightarrow [0,1]$ such that for every $S \subseteq \mathbb{R}^n$:

$$\mathbb{P}_{\boldsymbol{X}}(S) = \sum_{v \in S} P_{\boldsymbol{X}}(v)$$

If we know the joint probability distribution of a vector, we can also determine the probability distribution of each of its terms.

Proposition 2.5.13:

Given $\boldsymbol{X} = (X_1, \ldots, X_n)$, then for every relevant i: $\mathbb{P}_{X_i}(A) = \mathbb{P}_{\boldsymbol{X}}\left(\mathbb{R}^{i-1} \times A \times \mathbb{R}^{n-i}\right)$.

Proof:

Notice that:

$$\mathbb{P}_{\boldsymbol{X}}\left(\mathbb{R}^{i-1} \times A \times \mathbb{R}^{n-i}\right) = \mathbb{P}(X_1 \in \mathbb{R}, \ldots, X_{i-1} \in \mathbb{R}, X_i \in A, X_{i+1} \in \mathbb{R}, \ldots, X_n \in \mathbb{R})$$

And we know that $X_j \in \mathbb{R}$ is true, so this is just equal to:

$$\mathbb{P}(X_i \in A)$$

■

Notice that this means

$$\mathbb{P}(X=x) = \mathbb{P}(X=x, Y \in \mathbb{R}) = \sum_{y \in \mathbb{R}} \mathbb{P}(X=x, Y=y)$$

Definition 2.5.14:

If X is a random variable and A is an event with nonzero probability, then we define **conditional probability** on X given the event A by:

$$\mathbb{P}_{X|A}(S) = \mathbb{P}(X \in S \mid A) = \mathbb{P}_A(X \in S)$$

Definition 2.5.15:

Two random variables, X and Y, are **independent** if for every $A, B \subseteq \mathbb{R}$: $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$. This is denoted $X \perp\!\!\!\perp Y$ as usual.

Proposition 2.5.16:

X and Y are independent random variables if and only if for every $A \subseteq \mathbb{R}$ where $\mathbb{P}_X(A) > 0$: $\mathbb{P}_{Y|\{X \in A\}} = \mathbb{P}_Y$.

Proof:

Notice that:

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

And thus if X and Y are independent this equals $\mathbb{P}(Y \in B)$, so the distributions are the same. And if this is equal to $\mathbb{P}_Y(B)$, then we get $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$ for every A and B, which means X and Y are independent. ■

Proposition 2.5.17:

If X and Y are discrete random variables, then they are indpendent if and only if for every real x and y:

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

Proof:

($\Longrightarrow$) Let $A = \{x\}$ and $B = \{y\}$, then:

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

($\Longleftarrow$) Let $A, B \subseteq \mathbb{R}$, then:

$$\mathbb{P}(X \in A, Y \in B) = \sum_{(a,b) \in A \times B} \mathbb{P}(X = a, Y = b) = \sum_{a \in A} \sum_{b \in B} \mathbb{P}(X = a) \cdot \mathbb{P}(Y = b) =$$

$$\sum_{a \in A} \mathbb{P}(X = a) \cdot \sum_{b \in B} \mathbb{P}(Y = b) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$$

So we have that X and Y are independent.

■

Definition 2.5.18:

A set of random variables $\{X_i\}_{i=1}^n$ is **independent** if for every $E_1, \ldots, E_n \subseteq \mathbb{R}$:

$$\mathbb{P}(X_1 \in E_1, \ldots, X_n \in E_n) = \mathbb{P}(X_1, \in E_1) \cdots \mathbb{P}(X_n \in E_n)$$

And given an infinitely countable set of random variables $\{X_i\}_{i=1}^\infty$, they are independent if for every $n \in \mathbb{R}$, $\{X_i\}_{i=1}^n$ is independent.

Proposition 2.5.19:

If $\{X_i\}_{i\in I}$ is independent and I is coutable and $E_i \subseteq \mathbb{R}$ for every $i \in I$, then:

$$\mathbb{P}(\forall i \in I : X_i \in E_i) = \prod_{i\in I} \mathbb{P}(X_i \in E_i)$$

Proof:

If I is finite then this is true by definition. Otherwise we know that by **corollary 2.2.8**:

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty}\{X_i \in E_i\}\right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i \in E_i\}\right) = \lim_{n\to\infty} \prod_{i=1}^{n} \mathbb{P}(X_i \in E_i) = \prod_{i=1}^{\infty} \mathbb{P}(X_i \in E_i)$$

As required. ■

Definition 2.5.20:

A discrete random variable X has a **geometric distribution** over $p \neq 0$, denoted $X \sim \mathrm{Geo}(p)$ if for every $n \in \mathbb{N}_1$:

$$P_X(n) = (1-p)^{n-1} \cdot p$$

Note:

It is necessary to show that this is in fact a valid distribution. So we must show that $\sum_{n\in\mathbb{N}_1} \mathbb{P}(X = n) = 1$. If $p = 1$ this is true since $P_X(1) = p = 1$ and for every other n it is 0, so the sum is 1. Otherwise:

$$\sum_{n=1}^{\infty} \mathbb{P}(X = n) = p \cdot \sum_{n=1}^{\infty} (1-p)^{n-1} = p \cdot \sum_{n=0}^{\infty} (1-p)^n$$

This is a geometric sum (which converges since $(1-p) < 1$), so this is equal to:

$$p \cdot \frac{1}{1-(1-p)} = \frac{p}{p} = 1$$

As required.

Theorem 2.5.21:

If $\{X_i\}_{i=1}^{\infty}$ are independent random variables which have a distribution of $\mathrm{Ber}(p)$, then $\min\{k \in \mathbb{N}_1 \mid X_k = 1\}$ has a distribution of $\mathrm{Geo}(p)$.

The idea behind this theorem is that if you have a series of independent trials which can either succeed or fail with a probability of p, then the probability that you succeed for the first time on your nth try distributes geometrically.

Proof:

Let $Y = \min\{k \in \mathbb{N}_1 \mid X_k = 1\}$, we need to show $Y \sim \mathrm{Geo}(p)$. We know that $Y = n$ if and only if $X_1, \ldots, X_{n-1} = 0$ and $X_n = 1$, since Y tracks the *minimum* index where $X_k = 1$. So:

$$\mathbb{P}(Y = n) = \mathbb{P}(X_1 = 0, \ldots, X_{n-1} = 0, X_n = 1)$$

And since the X_is are independent this is equal to:

$$= \mathbb{P}(X_1 = 0) \cdots \mathbb{P}(X_{n-1} = 0) \cdot \mathbb{P}(X_n = 1) = (1-p) \cdots (1-p) \cdot p = (1-p)^{n-1} \cdot p$$

Which means Y distributes geometrically over p, as required.

■

Lemma 2.5.22:

Suppose X is a random variables whose support is $\mathbb{N}_1$, then $\mathbb{P}(X > n) = (1-p)^n$ for every $n \in \mathbb{N}_1$ if and only if $X \sim \text{Geo}(p)$.

Proof:

($\implies$) Notice that $\mathbb{P}(X = n) = \mathbb{P}(X > n-1) - \mathbb{P}(X > n)$ which is equal to in this case $(1-p)^{n-1} - (1-p)^n = (1-p)^{n-1} \cdot (1-(1-p)) = (1-p)^{n-1} \cdot p$ as required.

($\impliedby$) We will prove this through simple computation:

$$\mathbb{P}(X > n) = \sum_{x=n+1}^{\infty} \mathbb{P}(X = x) = p \cdot \sum_{x=n+1}^{\infty} (1-p)^{x-1} = p \cdot \frac{(1-p)^n}{p} = (1-p)^n$$

Since the sum is geometric.

■

Theorem 2.5.23 (Memorylessness of Geometric Distributions):

Suppose X is a random variables whose support is $\mathbb{N}_1$ and $\mathbb{P}(X = 1) < 1$. Then the following are equivalent:

(1) X distributes geometrically.

(2) $X \overset{d}{=} (X - 1 \mid X > 1)$

(3) For every $m \in \mathbb{N}_1$, $X \overset{d}{=} (X - m \mid X > m)$

The idea behind this theorem is that after m trials fail ($X > m$), the current state of the world is still the same as the beginning of the trials (distributively equivalent to X).

Proof:

(1) $\implies$ **(3)** Suppose $X \sim \text{Geo}(p)$. We will prove this through direct computation:

$$\mathbb{P}_{X-m|X>m}(n) = \mathbb{P}(X - n = m \mid X > m) = \frac{\mathbb{P}(X = n+m)}{\mathbb{P}(X > m)} = \frac{p(1-p)^{n+m-1}}{(1-p)^m} = p(1-p)^{n-1}$$

Which is a geometric distribution, as required.

(3) $\implies$ **(2)** This is trivial by letting $m = 1$.

(2) $\implies$ **(1)** Notice that $\mathbb{P}(X = n) = \mathbb{P}(X = n+1 \mid X > 1) = \frac{\mathbb{P}(X=n+1)}{\mathbb{P}(X>1)}$. So we get that

$$\mathbb{P}(X = n+1) = \mathbb{P}(X = n) \cdot \mathbb{P}(X > 1)$$

If we let $p := \mathbb{P}(X = 1)$, then $\mathbb{P}(X > 1) = 1 - p$ since the support of X is $\mathbb{N}_1$. So:

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

This is the definition of a geometric series (if we let $a_n = \mathbb{P}(X = n)$, we see that $a_{n+1} = (1-p)a_n$). So we get:

$$\mathbb{P}(X = n) = (1-p)^{n-1} \cdot \mathbb{P}(X = 1) = (1-p)^{n-1} \cdot p$$

As required.

■

Definition 2.5.24:

Let $n \in \mathbb{N}_1$ and $p \in [0,1]$. We say that a discrete random variable X has a **binomial distribution** over n and p, denoted $X \sim \text{Bin}(n,p)$ if for every natural k between 0 and n: $\mathbb{P}(X=k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$.

Note:

We must check that this is a valid distribution.

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

Which is equal to, by the binomial theorem:

$$= (p+1-p)^n = 1^n = 1$$

As required.

Theorem 2.5.25:

If $\{X_i\}_{i=1}^n$ are indpendent random variables such that $X_i \sim \text{Ber}(p)$, then $\sum_{i=1}^{n} X_i \sim \text{Bin}(n,p)$.

The idea behind this is that if you have n independent trials each with a probability of success of p, the probability that exactly k of them succeed has a binomial distribution.

Proof:

Let $Y := \sum_{i=1}^{n} X_i$. $Y = k$ if and only if k of the X_is are equal to 1 and the rest are 0. So we must choose a subset of size k of the X_is to be equal to 1, and there are $\binom{n}{k}$ choices for this. This means that:

$$\mathbb{P}(Y=k) = \sum_{I \in \mathcal{P}_k([n])} \mathbb{P}(\forall i \in I : X_i = 1, \forall i \notin I : X_i = 0)$$

Notice that for every I:

$$\mathbb{P}(\forall i \in I : X_i = 1, \forall i \notin I : X_i = 0) = \prod_{i \in I} \mathbb{P}(X_i = 1) \cdot \prod_{i \notin I} \mathbb{P}(X_i = 0)$$

Since the X_is are independent. This is equal to:

$$= \prod_{i \in I} p \cdot \prod_{i \notin I} (1-p) = p^{|I|} \cdot (1-p)^{n-|I|}$$

Since $I \in \mathcal{P}_k([n])$, $|I| = k$, so:

$$\mathbb{P}(Y=k) = \sum_{I \in \mathcal{P}_k([n])} p^k \cdot p^{n-k}$$

Since there are $\binom{n}{k}$ choices for I, this is equal to:

$$= \binom{n}{k} \cdot p^k \cdot p^{n-k}$$

As required. ■

Corollary 2.5.26:

If $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Bin}(m,p)$ and X and Y are independent, then $X+Y \sim \text{Bin}(n+m,p)$.

Proof:

Suppose $\{X_i\}_{i=1}^{n+m}$ are independent random variables with a distribution of $\text{Ber}(p)$. Then

$$X \overset{d}{=} \sum_{i=1}^{n} X_i \quad Y \overset{d}{=} \sum_{i=n+1}^{m} X_i$$

These are independent so $X+Y = \sum_{i=1}^{n+m} X_i$, which by the theorem above distributes $\text{Bin}(n+m,p)$, as required.

■

Definition 2.5.27:

Suppose $\{X_n\}_{n=1}^{\infty}$ is a set of random variables, and so is Y. These random variables don't need to necessarily be over the same probability space. We say that the **distributive limit** of X_n is Y, denoted $X_n \xrightarrow{d} Y$ if for every $A \subseteq \mathbb{R}$: $\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \mathbb{P}(Y \in A)$.
If X_n and Y both have a countable support S then this is equivalent to $\forall k \in S : \lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(Y = k)$.

Definition 2.5.28:

A discrete random variable X has a **Poisson Distribution** over $\lambda \in \mathbb{R}_{>0}$, denote $X \sim \text{Poi}(\lambda)$ if for every $n \in \mathbb{N}_0$, we have that $\mathbb{P}(X=n) = e^{-\lambda} \cdot \frac{\lambda^n}{n!}$.

Note:

This is a valid distribution since:

$$\sum_{n=0}^{\infty} \mathbb{P}(X=n) = e^{-\lambda} \cdot \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}$$

The right side is the powerseries expansion of e^{λ}, so this is equal to $e^{-\lambda} \cdot e^{\lambda} = 1$ as required.

Theorem 2.5.29:

Suppose $\{X_n\}_{n=1}^{\infty}$ is a series of random variables such that $X_n \sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$ for some positive real λ. Then

$$X_n \xrightarrow{d} \text{Poi}(\lambda)$$

Proof:

Let $k \in \mathbb{N}_0$. We sill show that $\lim_{n\to\infty} \mathbb{P}(X_n = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$. We know that:

$$\mathbb{P}(X_n = k) = \binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Now note that:

$$\binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k = \frac{n!}{n^k \cdot (n-k)!} \cdot \frac{\lambda^k}{k!} = \frac{n \cdot (n-1) \cdots (n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!}$$

Notice that the numerator and denominator of $\frac{n\cdot(n-1)\cdots(n-k+1)}{n^k}$ are both degree k polynomials with a leading coefficient

of 1, so the limit of this is 1. The limit of $\frac{\lambda^k}{k!}$ is obviously $\frac{\lambda^k}{k!}$ since it is independent of n. Next we have:

$$\left(1-\frac{\lambda}{n}\right)^{n-k}=\left(1-\frac{\lambda}{n}\right)^{n}\cdot\left(1-\frac{\lambda}{n}\right)^{-k}$$

The left has a limit of $e^{-\lambda}$, and the left has a limit of 1 (since the limit of $1-\frac{\lambda}{n}$'s limit is 1). So all in all the limit of $\mathbb{P}(X_n=k)$ is:

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

As required.

■

Proposition 2.5.30:

(1) If X and Y are independent random variables and $X\sim\text{Poi}(\lambda)$ and $Y\sim\text{Poi}(\mu)$ then $X+Y\sim\text{Poi}(\lambda+\mu)$.

(2) If $X\sim\text{Poi}(\lambda)$ and $Y\mid X=n\sim\text{Bin}(n,p)$ then $Y\sim\text{Poi}(p\lambda)$.

Proof:

(1) We will prove this through direct computation:

$$\mathbb{P}(X+Y=k)=\sum_{n=0}^{k}\mathbb{P}(X=n,Y=k-n)=\sum_{n=0}^{k}\mathbb{P}(X=n)\cdot\mathbb{P}(Y=k-n)=$$

$$=\sum_{n=0}^{k}e^{-\lambda}\cdot e^{-\mu}\cdot\frac{\lambda^n}{n!}\cdot\frac{\mu^{n-k}}{(k-n)!}=e^{-(\lambda+\mu)}\cdot\sum_{n=0}^{k}\frac{\lambda^n\cdot\mu^{n-k}}{n!\cdot(k-n)!}$$

Doing a bit of algebraic manipulation, this is equal to:

$$\frac{e^{-(\lambda+\mu)}}{k!}\cdot\sum_{n=0}^{k}\binom{k}{n}\lambda^n\cdot\mu^{n-k}=e^{-(\lambda+\mu)}\cdot\frac{(\lambda+\mu)^k}{k!}$$

As required.

(2) We know that since $Y\mid X=n\sim\text{Bin}(n,p)$, if $Y=k$ then $n\geq k$.

$$\mathbb{P}(Y=k)=\sum_{n=k}^{\infty}\mathbb{P}(Y=k\mid X=n)\cdot\mathbb{P}(X=n)=\sum_{n=k}^{\infty}\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}\cdot e^{-\lambda}\cdot\frac{\lambda^n}{n!}=$$

$$=e^{-\lambda}p^k\cdot\sum_{n=k}^{\infty}\binom{n}{k}\cdot\cdot(1-p)^{n-k}\cdot\frac{\lambda^n}{n!}$$

Notice that the term inside the sum is equal to:

$$\frac{\lambda^n\cdot(1-p)^{n-k}}{(n-k)!}$$

So the sum is equal to:

$$=\frac{e^{-\lambda}\cdot p^k\cdot\lambda^k}{k!}\cdot\sum_{n=0}^{\infty}\frac{\lambda^n\cdot(1-p)^n}{n!}$$

The sum is the powerseries of $e^{\lambda(1-p)}$, so this is equal to:

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

Which is the poisson distribution of $p\lambda$, as required.

Definition 2.5.31:

A discrete random variable X has a **hypergeomtric distribution** over $N, D, n \in \mathbb{N}_0$ if:

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

For relevant k. This is denoted $X \sim \mathrm{HG}(N, D, n)$.

Note:

Notice that if $X \sim \mathrm{HG}(N, D, n)$ then $k \leq D$, $n \leq N$, $k \leq n$, $D \leq N$, and $n - k \leq N - D$. So $0, D + n - N \leq k \leq D, n$.

Theorem 2.5.32:

Suppose an urn has N objects, D of which are considered "special". We remove (without repetition) n objects from the urn. Let X be the number of "special" objects removed, then $X \sim \mathrm{HG}(N, D, n)$.

Proof:

How many ways are there to remove k special objects out of n choices? Well first there are $\binom{D}{k}$ choices for which special objects to choose, and then there are another $\binom{N-D}{n-k}$ choices for the remaining objects (there are $N - D$ non-special objects, and $n - k$ objects which still need to be removed). So all in all there are $\binom{D}{K} \cdot \binom{N-D}{n-k}$ ways to choose k special objects out of n choices. There are $\binom{N}{n}$ ways to choose n objects, and since each choice is equally likely (since the probability of choosing any single object is equal), the probability that we choose k special objects is:

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

As required.

■

Note:

This also proves that hypergeometric distributions are valid probability distributions since they represent a valid probability situation. X must be equal to some k between $0, D + n - N$ and n, D, so the sum over all possible ks of $\mathbb{P}(X = k)$ must be 1.

2.6 Expected Values

Definition 2.6.1:

Given a random variable X, we define the **expected value** of X, denoted $\mathbb{E}[X]$ to be:

$$\mathbb{E}[X] := \sum_{x \in \mathbb{R}} x \cdot \mathbb{P}(X = x)$$

Since this is not a necessarily positive sum, and we are summing over a set $\mathbb{R}$ (so order doesn't matter), it is necessary for this sum to converge absolutely. So X has an expected value if and only if

$$\sum_{x \in \mathbb{R}} |x| \cdot \mathbb{P}(X = x) < \infty$$

Which is equivalent to $\mathbb{E}[|X|] < \infty$.

The expected value of a random variables gives a weighted average of the values of the random variable.

Theorem 2.6.2 (The Law of the Unconscious Statistician):

If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random vector of n random variables and $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ then

$$\mathbb{E}[f(\boldsymbol{X})] = \sum_{Y \in \mathbb{R}^n} f(Y) \cdot \mathbb{P}(\boldsymbol{X} = Y)$$

Proof:

We know that:

$$\mathbb{E}[f(\boldsymbol{X})] = \sum_{x \in \mathbb{R}} x \cdot \mathbb{P}(f(\boldsymbol{X}) = x)$$

But $f(\boldsymbol{X}(\omega)) = x$ if and only if $\boldsymbol{X}(\omega) \in f^{-1}\{x\}$. This means that

$$\mathbb{P}(f(\boldsymbol{X}) = x) = \mathbb{P}\left(\boldsymbol{X} \in f^{-1}\{x\}\right) = \sum_{Y \in f^{-1}\{x\}} \mathbb{P}(\boldsymbol{X} = Y)$$

So we have that

$$\mathbb{E}[f(\boldsymbol{X})] = \sum_{x \in \mathbb{R}} x \cdot \sum_{Y \in f^{-1}\{x\}} \mathbb{P}(\boldsymbol{X} = Y) = \sum_{x \in \mathbb{R}} \sum_{Y \in f^{-1}\{x\}} f(Y) \cdot \mathbb{P}(\boldsymbol{X} = Y)$$

Note that summing over every $x \in \mathbb{R}$ and $Y \in f^{-1}\{x\}$ is equal to summing over every $Y \in \mathbb{R}^n$, since for every real vector Y, there is exactly one x where $Y \in f^{-1}\{x\}$, and that is $f(Y)$. So this is equal to

$$\sum_{Y \in \mathbb{R}^n} f(Y) \cdot \mathbb{P}(\boldsymbol{X} = Y)$$

As required.

■

In the special case where $n = 1$, so $\boldsymbol{X} = (X)$, and $f : \mathbb{R} \longrightarrow \mathbb{R}$, we have that:

$$\mathbb{E}[f(X)] = \sum_{x \in \mathbb{R}} f(x) \cdot \mathbb{P}(X = x)$$

Theorem 2.6.3:

The following are true:

(1) If $X \overset{as}{\geq} 0$ then $\mathbb{E}[X] \geq 0$ (this means $\mathbb{P}(X \geq 0) = 1$).

(2) If $X \overset{d}{=} Y$ then $\mathbb{E}[X] = \mathbb{E}[Y]$.

(3) $\mathbb{E}$ is linear: $\mathbb{E}[\alpha X + \beta Y] = \alpha\,\mathbb{E}[X] + \beta\,\mathbb{E}[Y]$.

(4) $\mathbb{E}$ is monotonic: If $X \overset{as}{\geq} Y$ then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

(5) If X and Y are indpendent, $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

(6) If X has a support in $\mathbb{N}$, then $\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n)$.

Proof:

(1) Since $\mathbb{P}(X \geq 0) = 1$, $\mathbb{P}(X < 0) = 0$, so for every $x < 0$, $\mathbb{P}(X = x) = 0$. Therefore:

$$\mathbb{E}[X] = \sum_{x>0} x \cdot \mathbb{P}(X = x)$$

Which is a positive sum.

(2) Since both random variables have the same distribution, the sum of $x\mathbb{P}(X = x)$ is equal to $x\mathbb{P}(Y = x)$, so the expected values are equal.

(3) First let us show that $\alpha X + \beta Y$ has an expected value. So we need to show that $\mathbb{E}\left[\left|\alpha X + \beta Y\right|\right] < \infty$. But notice by **The Law of the Unconscious Statistician**:

$$\mathbb{E}\left[\left|\alpha X + \beta Y\right|\right] = \sum_{x,y\in\mathbb{R}} \left|\alpha x + \beta y\right| \cdot \mathbb{P}(X = x, Y = y) \leq$$

$$\leq |\alpha| \sum_{x\in\mathbb{R}} |x| \cdot \sum_{y\in\mathbb{R}} \mathbb{P}(X = x, Y = y) + \left|\beta\right| \sum_{y\in\mathbb{R}} \left|y\right| \cdot \sum_{x\in\mathbb{R}} \mathbb{P}(X = x, Y = y)$$

Notice that:

$$|\alpha| \sum_{x\in\mathbb{R}} |x| \cdot \sum_{y\in\mathbb{R}} \mathbb{P}(X = x, Y = y) = |\alpha| \sum_{x\in\mathbb{R}} |x| \cdot \mathbb{P}(X = x)$$

Which converges since $\mathbb{E}[X]$ exists. Similar for the other term. So $\mathbb{E}\left[\left|\alpha X + \beta Y\right|\right] < \infty$, as required.
And by **The Law of the Unconscious Statistician** again:

$$\mathbb{E}[\alpha X + \beta Y] = \sum_{x,y\in\mathbb{R}} (\alpha x + \beta y) \cdot \mathbb{P}(X = x, Y = y)$$

Doing a very similar process to the one above, we see that this is equal to:

$$\alpha \sum_{x\in\mathbb{R}} x \cdot \sum_{y\in\mathbb{R}} \mathbb{P}(X = x, Y = y) + \beta \sum_{y\in\mathbb{R}} y \cdot \sum_{x\in\mathbb{R}} \mathbb{P}(X = x, Y = y)$$

Which is equal to

$$\alpha \sum_{x\in\mathbb{R}} x \cdot \mathbb{P}(X = x) + \beta \sum_{y\in\mathbb{R}} y \cdot \mathbb{P}(Y = y) = \alpha\,\mathbb{E}[X] + \beta\,\mathbb{E}[Y]$$

As required.

(4) This means $X - Y \overset{as}{\geq} 0$, so $\mathbb{E}[X - Y] \geq 0$, so $\mathbb{E}[X] - \mathbb{E}[Y] \geq 0$, and therefore $\mathbb{E}[X] \geq \mathbb{E}[Y]$, as required.

(5) Assuming the expected value exists, we see that:

$$\mathbb{E}[X \cdot Y] = \sum_{x,y\in\mathbb{R}} xy\,\mathbb{P}(X = x, Y = y) = \sum x \in \mathbb{R} x\,\mathbb{P}(X = x) \cdot \sum_{y\in\mathbb{R}} y\,\mathbb{P}(Y = y) = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Since $|X|$ and $|Y|$ are also independent, $\mathbb{E}[|X \cdot Y|] = \mathbb{E}[|X|] \cdot \mathbb{E}[|Y|]$ by above, and these both converge since X and Y have expected value, so $\mathbb{E}[|X \cdot Y|] < \infty$ as required.

(6) Notice that $n \cdot \mathbb{P}(X = n) = \sum_{k=1}^{n} \mathbb{P}(X = n)$, so:

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} n\mathbb{P}(X = n) = \sum_{n=1}^{\infty}\sum_{k=1}^{n} \mathbb{P}(X = n)$$

This is summing over $n \in \mathbb{N}$, $n \geq k$, so we can reverse the order of summation to get:

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

As required.

■

Proposition 2.6.4:

(1) $\mathbb{E}[\text{Ber}(p)] = p$

(2) $\mathbb{E}[\text{Bin}(n,p)] = np$

(3) $\mathbb{E}[\text{Geo}(p)] = \frac{1}{p}$

(4) $\mathbb{E}[\text{Unif}[a,b]] = \frac{a+b}{2}$

(5) $\mathbb{E}[\text{Poi}(\lambda)] = \lambda$

Proof:

(1) With some direct computation, we see that $\mathbb{E}[\text{Ber}(p)] = 1 \cdot p + 0 \cdot (1-p) = p$, as required.

(2) Since $\text{Bin}(n,p) = \text{Ber}(p) + \cdots + \text{Ber}(p)$, $\mathbb{E}[\text{Bin}(n,p)] = \mathbb{E}[\text{Ber}(p)] + \cdots + \mathbb{E}[\text{Ber}(p)] = p + \cdots + p = np$, as required.

(3) Notice that if $p < 1$:

$$\mathbb{E}[\text{Geo}(p)] = p \cdot \sum_{n=1}^{\infty} n \cdot (1-p)^{n-1}$$

Notice that

$$-\frac{d}{dp}\sum_{n=1}^{\infty}(1-p)^n = \sum_{n=1}^{\infty} n(1-p)^{n-1}$$

This is true for $p < 1$ since the radius of convergence of this powerseries is 1. And we know that:

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

Whose derivative is $-\frac{1}{p^2}$. So:

$$\sum_{n=1}^{\infty} n(1-p)^{n-1} = \frac{1}{p^2}$$

And therefore $\mathbb{E}[\text{Geo}(p)] = \frac{1}{p}$.

If $p = 1$ then the sum is just equal to 1, since for $n > 1$ $(1-p)^{n-1} = 0$.

(4)

$$\mathbb{E}[\text{Unif}[a,b]] = \sum_{n=a}^{b} n \cdot \frac{1}{b-a+1} = \frac{1}{b-a+1} \cdot \frac{b-a+1}{2}(a+b) = \frac{a+b}{2}$$

(5)

$$\mathbb{E}[\text{Poi}(\lambda)] = e^{-\lambda} \cdot \sum_{n=0}^{\infty} n \cdot \frac{\lambda^n}{n!} = e^{-\lambda} \cdot \sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!} = \lambda e^{-\lambda} \cdot \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = \lambda$$

As required.

■

2.7 Variance

We can think of expected values as an approximation of a random variable, and then a good idea is to come up with a measure for how good of an approximation this is. We call this approximation the random variable's *variance*.

Definition 2.7.1:

Given a random variable X, we define its **variance** to be:

$$\operatorname{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

This is of course assuming that it exists.

Note that we can't just define it to be $\mathbb{E}[X - \mathbb{E}[X]]$ which would be more natural, since this equals $\mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$. Notice that the variance is equal to:

$$= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] = \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

Theorem 2.7.2:

The following are true:

(1) $\operatorname{Var}(X) \geq 0$.

(2) $\operatorname{Var}(X) = 0$ if and only if there exists some $c \in \mathbb{R}$ such that $X \overset{as}{=} c$.

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

(4) $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$.

(5) If X and Y are independent, then $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.

Proof:

(1) Since $(X - \mathbb{E}[X])^2 \geq 0$, so is its expected value, which is the variance of X.

(2) Note that if $X \overset{as}{=} c$, then $\mathbb{E}[c] = c = \mathbb{E}[X]$. So if $\operatorname{Var}(X) = 0$, notice that this means:

$$0 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_{k \in \mathbb{R}} k^2 \, \mathbb{P}(X - \mathbb{E}[X] = k)$$

And $k^2 \geq 0$, for $k \neq 0$ the term must be 0, so $\mathbb{P}(X - \mathbb{E}[X] = k) = 0$ for $k \neq 0$. And since this is a probability function, this means that $\mathbb{P}(X - \mathbb{E}[X] = 0) = 1$, so by definition $X \overset{as}{=} \mathbb{E}[X]$. For the converse, notice $X \overset{d}{=} \mathbb{E}[X]$, so $\operatorname{Var}(X) = \mathbb{E}\left[(\mathbb{E}[X] - \mathbb{E}[X])^2\right] = \mathbb{E}[0] = 0$.

(3) Notice that

$$\operatorname{Var}(X + a) = \mathbb{E}\left[(X + a - \mathbb{E}[X + a])^2\right] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \operatorname{Var}(X)$$

(4) Notice that

$$\operatorname{Var}(aX) = \mathbb{E}\left[(aX - \mathbb{E}[aX])^2\right] = \mathbb{E}\left[a^2(X - \mathbb{E}[X])^2\right] = a^2 \operatorname{Var}(X)$$

(5) Plugging in $X + Y$ to the formula for variance we found above gives:

$$\operatorname{Var}(X + Y) = \mathbb{E}\left[(X + Y)^2\right] - \mathbb{E}[X + Y]^2 = \mathbb{E}\left[X^2\right] + 2\mathbb{E}[XY] + \mathbb{E}\left[Y^2\right] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2$$

Since X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, so this is equal to:

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

As required.

■

Now let's compute the variance of some distributions:

Proposition 2.7.3:

(1) $\text{Var}(\text{Ber}(p)) = p - p^2$

(2) $\text{Var}(\text{Bin}(n,p)) = n(p-p^2)$

(3) $\text{Var}(\text{Unif}[n]) = \frac{n^2-1}{12}$

(4) $\text{Var}(\text{Poi}(\lambda)) = \lambda$

(5) $\text{Var}(\text{Geo}(p)) = \frac{1-p}{p^2}$

Proof:

(1) Notice that $X^2 \text{Ber}(p)$ since it is actually equal to X. So $\text{Var}(X) = \mathbb{E}[X] - \mathbb{E}[X]^2 = p - p^2$.

(2) Since $\text{Bin}(n,p)$ is distributively equal to the sum of n independent bernoulli distributions, and the variance of the sum of independent random variables is equal to the sum of the variances, this is equal to $n(p-p^2)$ as required.

(3) Notice that:

$$\mathbb{E}\left[X^2\right] = \sum_{k=1}^{n} k^2 \cdot \mathbb{P}(X=k) = \frac{1}{n} \cdot \sum_{k=1}^{n} k^2 = \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$

And so:

$$\text{Var}(X) = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}$$

And notice that $\text{Unif}[a,b] = \text{Unif}[b-a+1] + a - 1$, so:

$$\text{Var}(\text{Unif}[a,b]) = \frac{b^2 + a^2 - 2ab + 2b - 2a}{12}$$

(4) Notice that:

$$\mathbb{E}\left[X^2\right] = e^{-\lambda} \cdot \sum_{k=0}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \cdot \sum_{k=0}^{\infty} (k+1)\frac{\lambda^k}{k!} = \lambda e^{-\lambda}(e^{\lambda} + \lambda e^{\lambda}) = \lambda + \lambda^2$$

And so the variance is equal to $\lambda + \lambda^2 - \lambda^2 = \lambda$.

(5) This is left as an exercise (or look it up). The computation doesn't really add anything of value to the discussion.

■

Proposition 2.7.4:

Suppose X is a random variable which has variance. Then

$$\text{Var}(X) = \min_{a \in \mathbb{R}} \mathbb{E}\left[(X-a)^2\right]$$

So the minimum is when $a = \mathbb{E}[X]$.

Proof:

Let $Y = X - \mathbb{E}[X]$ and let $\varepsilon \neq 0$. We will show that $\mathbb{E}\left[(Y+\varepsilon)^2\right] > \text{Var}(X) = \text{Var}(Y)$. Furthermore, since $\mathbb{E}[Y] = 0$, $\mathbb{E}[Y+\varepsilon] = \varepsilon$, so:

$$\mathbb{E}\left[(Y+\varepsilon)^2\right] = \text{Var}(Y+\varepsilon) + \mathbb{E}[Y+\varepsilon]^2 = \text{Var}(Y+\varepsilon) + \varepsilon^2 = \text{Var}(X) + \varepsilon^2$$

So $\mathbb{E}\left[(Y+\varepsilon)^2\right] > \text{Var}(X)$ as required (since $\varepsilon^2 > 0$).

■

Notice that if X and Y are two random variables, we showed above that (this is a trivial result from the definition of variance):

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y])$$

This rightmost term turns out to be somewhat important, and it offers a sort of measure as to how dependent two random variables are. So like what we do with every significant mathematical object, we'll give it a name.

Definition 2.7.5:

Given two random variables X and Y, their **covariance** is:

$$\operatorname{Cov}(X,Y) := \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Note that this is equal to $\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$. As we remarked above, $\operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y + 2\operatorname{Cov}(X,Y)$. So it follows then that if X and Y are independent, $\operatorname{Cov}(X,Y) = 0$. And $\operatorname{Cov}(X,X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \operatorname{Var}(X)$.

Theorem 2.7.6:

(1) Covariance is symmetric: $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$.

(2) $\operatorname{Cov}(\alpha X + \beta Y, Z) = \alpha \operatorname{Cov}(X,Z) + \beta \operatorname{Cov}(Y,Z)$.

(3) $\operatorname{Cov}(X+\alpha, Y) = \operatorname{Cov}(X,Y)$

Proof:

(1) This is trivial by the definition of covariance.

(2) By definition:

$$\operatorname{Cov}(\alpha X + \beta Y, Z) = \mathbb{E}[(\alpha X + \beta Y)Z] - \mathbb{E}[\alpha X + \beta Y]\mathbb{E}[Z] =$$

$$= \alpha\mathbb{E}[XZ] + \beta\mathbb{E}[YZ] - \alpha\mathbb{E}[X]\mathbb{E}[Z] - \beta\mathbb{E}[Y]\mathbb{E}[Z] = \alpha\operatorname{Cov}(X,Z) + \beta\operatorname{Cov}(Y,Z)$$

As required.

(3) Notice that $\operatorname{Cov}(\alpha, Y) = \mathbb{E}[\alpha Y] - \mathbb{E}[\alpha]\mathbb{E}[Y] = \alpha\mathbb{E}[Y] - \alpha\mathbb{E}[Y] = 0$, and $\operatorname{Cov}(X+\alpha, Y) = \operatorname{Cov}(X,Y) + \operatorname{Cov}(\alpha, Y) = \operatorname{Cov}(X,Y)$ as required.

■

Theorem 2.7.7:

If $\{X_i\}_{i=1}^n$ are random variables then:

$$\operatorname{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \operatorname{Var}(X_i) + 2\sum_{1\le i<j\le n} \operatorname{Cov}\left(X_i, X_j\right)$$

Proof:

We can show this through induction. The trivial case of $n = 1$ is trivial, and the other base case $n = 2$ was shown above. For the inductive step, notice that in the sum $\sum_{i=1}^{n+1} X_i$, we can treat $X_n + X_{n+1}$ as one term and get a sum of n random variables. So:

$$\operatorname{Var}\left(\sum_{i=1}^{n+1} X_i\right) = \sum_{i=1}^{n-1} \operatorname{Var}(X_i) + \operatorname{Var}(X_n + X_{n+1}) + 2\sum_{1\le i<j\le n-1} \operatorname{Cov}\left(X_i, X_j\right) + 2\sum_{i=1}^{n-1} \operatorname{Cov}(X_i, X_n + X_{n+1})$$

Since $\operatorname{Var}(X_n + X_{n+1}) = \operatorname{Var}(X_n) + \operatorname{Var}(X_{n+1}) + 2\operatorname{Cov}(X_n, X_{n+1})$ and $\operatorname{Cov}(X_i, X_n + X_{n+1}) = \operatorname{Cov}(X_i, X_n) + \operatorname{Cov}(X_i, X_{n+1})$,

this is equal to:

$$= \sum_{i=1}^{n+1} \text{Var}(X_i) + 2\,\text{Cov}(X_n, X_{n+1}) + 2 \sum_{1 \le i < j \le n} \text{Cov}\left(X_i, X_j\right) + 2 \sum_{i=1}^{n-1} \text{Cov}(X_i, X_{n+1})$$

Since for $i = n$, $\text{Cov}(X_i, X_{n+1}) = \text{Cov}(X_n, X_{n+1})$, this is equal to:

$$= \sum_{i=1}^{n+1} \text{Var}(X_i) + 2 \sum_{1 \le i < j \le n+1} \text{Cov}\left(X_i, X_j\right)$$

As required.

■

An alternative way of writing this is:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{1 \le i \ne j \le n} \text{Cov}\left(X_i, X_j\right)$$

Since for every two indexes a and b, the term $\text{Cov}(X_a, X_b)$ will be added twice, since $\text{Cov}(X_b, X_a) = \text{Cov}(X_a, X_b)$ will also be added. And since $\text{Var}(X_i) = \text{Cov}(X_i, X_i)$ we can rewrite this again as:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{1 \le i, j \le n} \text{Cov}\left(X_i, X_j\right)$$

2.8 Approximations and Bounds

So now that we've discussed expected values and variance, what exactly are they useful for? Recall what we're trying to study in probability theory: probability. It turns out that variance and expected values, along with being interesting on their own, also provide useful tools for studying probability. A big contribution of theirs is the many bounds and approximations they provide for probability. This is best demonstrated in the following theorems.

Theorem 2.8.1 (Markov's Inequality):

Suppose $X \overset{as}{\geq} 0$ and X has an expected value. Then for any positive real a:

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

Proof:

We know that $X \overset{as}{\geq} a \cdot \mathbb{1}_{\{X \geq a\}}$ since if $X \geq a$ then $a \cdot \mathbb{1}_{\{X \geq a\}} = a$, and if $0 \leq X < a$ then it is equal to 0. So this means:

$$\mathbb{E}[X] \geq \mathbb{E}\left[a \cdot \mathbb{1}_{\{X \geq a\}}\right] = a \cdot \mathbb{E}\left[\mathbb{1}_{\{X \geq a\}}\right]$$

Now recall that an indicator function has an expected value of the probability of the event it indicates (since it has a bernoulli distribution over this parameter). So:

$$\mathbb{E}[X] \geq a \cdot \mathbb{P}(X \geq a)$$

Which means

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

As required. ■

Another way we can write Markov's inequality is by:

$$\mathbb{P}(X \geq b \cdot \mathbb{E}[X]) \leq \frac{1}{b}$$

For $b > 0$.

Exercise:

Suppose $\{X_i\}_{i=1}^n$ is a series of independent random variables such that $X_i \sim \text{Unif}[N]$. Find an upper bound for the probability that there are at least ℓ collisions between the random variables.

Solution:

First, let us define indicator functions which indicate whether or not two random variables are equal:

$$Y_{i,j} = \mathbb{1}_{\{X_i = X_j\}}$$

And we'll let Y be equal to the total number of collisions, which is the sum of all $Y_{i,j}$s for $i < j$:

$$Y = \sum_{i<j} Y_{i,j}$$

Since $Y_{i,j} \overset{as}{\geq} 0$, $Y \overset{as}{\geq} 0$. And since expected values are linear:

$$\mathbb{E}[Y] = \sum_{i<j} \mathbb{E}\left[Y_{i,j}\right] = \sum_{i<j} \mathbb{P}\left(X_i = X_j\right)$$

The probability that $X_i = X_j$ is equal to $\frac{1}{N}$ since if we set X_i's value, the probability that X_j is equal to that value is $\frac{1}{N}$.

Since there are $\binom{n}{2}$ values for i, j such that $i < j$ (take any 2-length subset of $[n]$ which naturally gives i, j).

$$\mathbb{E}[Y] = \sum_{i<j} \frac{1}{N} = \binom{n}{2} \cdot \frac{1}{N}$$

So all in all:

$$\mathbb{P}(Y \geq \ell) \leq \frac{\mathbb{E}[Y]}{\ell} = \binom{n}{2} \cdot \frac{1}{N \cdot \ell}$$

Exercise (The Coupon Collector's Problem):

Suppose there are n types of coupons, and you keep collecting coupons until you have all n types. Further suppose that the probability of collecting any type of coupon is equal (and thus distributes uniformly over $[n]$).

(1) How many coupons must be collected in order to have a probability of having all types of coupons with a probability $\geq \frac{1}{c}$?

(2) Show that the probability of collecting more than $2n\log(n)$ coupons (without getting all types) is $\leq \frac{1}{n}$.

Solution:

(1) Let Y_k be the minimum number of coupons collected in order to get k distinct coupons. So we want to analyze Y_n. Notice that $Y_0 = 0$ and:

$$Y_n = Y_n - Y_0 = \sum_{k=1}^{n} Y_k - Y_{k-1}$$

And further notice that the probability that $\mathbb{P}(Y_k - Y_{k-1} = t)$ is equal to $\frac{n-(k-1)}{n} \cdot \left(\frac{k-1}{n}\right)^{t-1}$ since $\frac{n-(k-1)}{n}$ is the probability we choose a different type coupon on the tth attempt after Y_{k-1}, and $\left(\frac{k-1}{n}\right)^{t-1}$ is the probability that we choose one of the $k-1$ types of coupons for the $t-1$ attempts before that. This means that $Y_k - Y_{k-1} \sim \text{Geo}\left(\frac{n-k+1}{n}\right)$. This means that:

$$\mathbb{E}[Y_n] = \sum_{k=1}^{n} \mathbb{E}[Y_k - Y_{k-1}] = \sum_{k=1}^{n} \frac{n}{n-k+1}$$

This is just the sum from the opposite direction of:

$$= n \cdot \sum_{k=1}^{n} \frac{1}{k}$$

And we know that

$$\sum_{k=1}^{n} \frac{1}{k} \leq \log(n) + 1$$

So all in all we have:

$$\mathbb{P}(Y_n \geq \ell) \leq \frac{\mathbb{E}[Y_n]}{\ell} \leq \frac{n}{\ell}(\log(n) + 1)$$

So if we set $\ell 2n(\log(n) + 1)$, we get that

$$\mathbb{P}(Y_n \geq \ell) \leq \frac{1}{2}$$

Which means that $\mathbb{P}(Y_n < \ell) > \frac{1}{2}$. So if we collect $2n(\log(n) + 1)$ coupons, we have a probability of success of greater than $\frac{1}{2}$.

(2) Let A_j^k denote the event where we collect k coupons but don't get the jth coupon. Thus $\mathbb{P}\left(A_j^k\right) = \left(1 - \frac{1}{n}\right)^k$ since the probability of not collecting the jth coupon each time is $1 - \frac{1}{n}$. The event that we collect k coupons and we

don't get one of the types is the union (relative to j) of A_j^k. And by the union bound we get that:

$$\mathbb{P}\left(\bigcup_{j=1}^{n} A_j^k\right) \leq \sum_{j=1}^{n} \mathbb{P}\left(A_j^k\right) \leq \sum_{j=1}^{n} \left(1-\frac{1}{n}\right)^k = n \cdot \left(1-\frac{1}{n}\right)^k$$

Notice that $\left(1-\frac{1}{n}\right)^k = \left(1-\frac{1}{n}\right)^{n\cdot\frac{k}{n}} \leq e^{-\frac{k}{n}}$. So if $k = 2n\log(n)$, we get that the probability is less than:

$$\leq n \cdot e^{-2\log(n)} = n \cdot n^{-2} = \frac{1}{n}$$

As required.

Theorem 2.8.2 (Chebyshev's Inequality):

If X is a random variable with variance, then for every positive real a:

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

Proof:

Let $Y := (X - \mathbb{E}[X])^2$, so $Y \geq 0$ and Y has expected value since $\mathbb{E}[Y] = \mathrm{Var}(X)$. Notice that by **Markov's Inequality** we have that for every $b > 0$:

$$\mathbb{P}(Y \geq b) \leq \frac{\mathbb{E}[Y]}{b} = \frac{\mathrm{Var}(X)}{b}$$

For $a > 0$ notice that $Y \geq a^2$ is equal to $(X - \mathbb{E}[X])^2 \geq a^2 = |X - \mathbb{E}[X]| \geq a$, so:

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

As required.

■

Notice that:

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

Theorem 2.8.3 (The Weak Law of Large Numbers):

Suppose $\{X_i\}_{i=1}^{\infty}$ is a series of random variables which all have the same distribution and variance (therefore their expected values and variances are all equal as well). Let X be a random variable which represents their distribution (ie. $X_i \overset{d}{=} X$ for every i). Then for every $\varepsilon > 0$ we have that:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbb{E}[X]\right| > \varepsilon\right) = 0$$

What this means is that the average result of the random variables does not diverge much from the expected value of the random variables.

Proof:

Let $\overline{X}_n := \frac{1}{n}\sum_{k=1}^{n} X_k$. Therefore:

$$\mathbb{E}\left[\overline{X}_n\right] = \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{E}[X_k] = \frac{1}{n} \cdot n \cdot \mathbb{E}[X] = \mathbb{E}[X]$$

And since the random variables are all independent we have that:

$$\operatorname{Var}\left(\overline{X}_n\right) = \frac{1}{n^2} \cdot \sum_{k=1}^{n} \operatorname{Var}(X_k) = \frac{1}{n^2} \cdot n \cdot \operatorname{Var}(X) = \frac{\operatorname{Var}(X)}{n}$$

By **Chebyshev's Inequality** we have that:

$$\mathbb{P}\left(\left|\frac{1}{n} \cdot \sum_{k=1}^{n} - \mathbb{E}[X]\right| > \varepsilon\right) = \mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]\right| > \varepsilon\right) \leq \frac{\operatorname{Var}\left(\overline{X}_n\right)}{\varepsilon^2} = \frac{1}{n} \cdot \frac{\operatorname{Var}(X)}{\varepsilon^2}$$

Since $\frac{\operatorname{Var}(X)}{\varepsilon^2}$ is constant

$$\lim_{n\to\infty} \frac{1}{n} \cdot \frac{\operatorname{Var}(X)}{\varepsilon^2} = 0$$

And since the probability is non-negative, by the squeeze theorem we have that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^{n} X_k - \mathbb{E}[X]\right| > \varepsilon\right) = 0$$

As required.

■

2.9 Conditional Expectation

Definition 2.9.1:

If X is a discrete random variable with an expected value and A is an event such that $\mathbb{P}(A) > 0$, we define the **conditional expectation** of X relative to A by:

$$\mathbb{E}[X \mid A] \coloneqq \sum_{x \in \mathbb{R}} x \cdot \mathbb{P}(X = x \mid A)$$

This definition begs the question, does conditional expectation always exist if $\mathbb{E}[X]$ does? The answer is yes.

Proposition 2.9.2:

Conditional expectation is well-defined in the regard stated above.

Proof:

Notice that by definition:

$$\mathbb{E}[X \mid A] = \sum_{x \in \mathbb{R}} x \cdot \frac{\mathbb{P}(X = x, A)}{\mathbb{P}(A)} = \frac{1}{\mathbb{P}(A)} \cdot \sum_{x \in \mathbb{R}} x \cdot \mathbb{P}(X = x, A)$$

Now note that that sum must converge since $\mathbb{P}(X = x, A) \leq \mathbb{P}(X = x)$ so absolutely the sum is less than the expected value of X, and converges. But we can squeeze out a bit more from this. Notice that if we instead change our point of view from events to random variables, the event A occurring is the same as $\mathbb{1}_A$ being 1. So $\mathbb{P}(X = x, A) = \mathbb{P}(X = x, \mathbb{1}_A = 1)$. Furthermore, notice that $\mathbb{1}_A \cdot X = x$ if and only if $\mathbb{1}_A = 1$ and $X = x$ or $x = 0$. But since $x = 0$ doesn't contribute to the sum, we get that:

$$\mathbb{E}[X \mid A] = \frac{1}{\mathbb{P}(A)} \cdot \sum_{x \in \mathbb{R}} x \cdot \mathbb{P}(\mathbb{1}_A \cdot X = x) = \frac{1}{\mathbb{P}(A)} \cdot \mathbb{E}[\mathbb{1}_A \cdot X]$$

And since $\mathbb{1}_A \cdot X \overset{as}{\leq} X$, this converges.

■

Similar to the **Law of Total Probability Version Two**, we have a similar situation with expectation and conditional expectation:

Proposition 2.9.3:

If X is a discrete random variable and $\{A_i\}_{i \in I}$ is a partition of Ω

$$\mathbb{E}[X] = \sum_{i \in I} \mathbb{E}[X \mid A_i] \cdot \mathbb{P}(A_i)$$

Proof:

This is simple and can be proven with some simple algebraic manipulation:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{R}} x \cdot \mathbb{P}(X = x)$$

Which is equal to by **Law of Total Probability Version Two**:

$$= \sum_{x \in \mathbb{R}} x \cdot \sum_{i \in I} \mathbb{P}(X = x \mid A_i) \cdot \mathbb{P}(A_i) = \sum_{i \in I} \mathbb{P}(A_i) \cdot \sum_{x \in \mathbb{R}} x \cdot \mathbb{P}(X = x \mid A_i) = \sum_{i \in I} \mathbb{E}[X \mid A_i] \cdot \mathbb{P}(A_i)$$

As required.

■

Proposition 2.9.4:

If N is a discrete random variable with a natural support and expected value, and $\{X_i\}_{i=1}^{\infty}$ is a sequence of random variables which have the same distribution (suppose $X_i \overset{d}{=} X$), then:

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[X] \cdot \mathbb{E}[N]$$

Proof:

We can use the previous proposition to see that:

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{n=1}^{\infty} \mathbb{E}\left[\sum_{i=1}^{N} X_i \,\middle|\, N = n\right] \cdot \mathbb{P}(N = n)$$

Notice that

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i \,\middle|\, N = n\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X]_i = n \cdot \mathbb{E}[X]$$

So:

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[X] \cdot \sum_{n=1}^{\infty} n \cdot \mathbb{E}[N] = \mathbb{E}[X] \cdot \mathbb{E}[N]$$

■

Definition 2.9.5:

Suppose X is a random variable with expected value, and Y is a random variable over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then we define a random variable:

$$\mathbb{E}[X \mid Y] : \Omega \longrightarrow \mathbb{R}$$

Such that for every $\omega \in \Omega$:

$$\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid Y = Y(\omega)]$$

If $\mathbb{P}(Y = Y(\omega))$ is non-zero, and we can define it to be 0 otherwise.

Conditional expectation is very useful when tring to compute expected values and variance for random variables which are very dependent on another. This is best demonstrated by the following theorems:

Theorem 2.9.6 (Law of Total Expectation):

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

Proof:

Let's start by introducing a new random variable, $\mathbb{P}(X = x \mid Y)$ for every x such that

$$\mathbb{P}(X = x \mid Y)(\omega) = \mathbb{P}(X = x \mid Y = Y(\omega))$$

Now it's quite simple to see that

$$\mathbb{E}[X \mid Y] = \sum_{x \in \mathbb{R}} x \cdot \mathbb{P}(X = x \mid Y)$$

As passing ω to the right hand side gives us precisely the definition of $\mathbb{E}[X \mid Y = Y(\omega)]$.
Now using this we see:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}\left[\sum_{x \in \mathbb{R}} x \cdot \mathbb{P}(X = x \mid Y)\right]$$

And applying the result from an above proposition, this is equal to:

$$\sum_{y\in\mathbb{R}}\mathbb{E}\left[\sum_{x\in\mathbb{R}}x\cdot\mathbb{P}(X=x\,\middle|\,Y=y)\right]\cdot\mathbb{P}(Y=y)$$

The expected value is just a constant so this is equal to:

$$\sum_{y\in\mathbb{R}}\sum_{x\in\mathbb{R}}x\cdot\mathbb{P}(X=x\,\middle|\,Y=y)\cdot\mathbb{P}(Y=y)=\sum_{x\in\mathbb{R}}x\cdot\sum_{y\in\mathbb{R}}\mathbb{P}(X=x,Y=y)=\sum_{x\in\mathbb{R}}x\cdot\mathbb{P}(X=x)=\mathbb{E}[X]$$

As required.

■

Proposition 2.9.7:

(1) If X and Y are independent then $\mathbb{E}[X\mid Y]=\mathbb{E}[X]$.

(2) If f is a real function then $\mathbb{E}[f(Y)\cdot X\mid Y]=f(Y)\cdot\mathbb{E}[X\mid Y]$.

Proof:

(1) Let $\omega\in\Omega$, then

$$\mathbb{E}[X\mid Y](\omega)=\mathbb{E}[X\mid Y=Y(\omega)]=\sum_{x\in\mathbb{R}}x\cdot\mathbb{P}(X=x\mid Y=Y(\omega))=\sum_{x\in\mathbb{R}}x\cdot\mathbb{P}(X=x)=\mathbb{E}[X]$$

(2) Let $\omega\in\Omega$ then:

$$\mathbb{E}[f(Y)\cdot X\mid Y]=\mathbb{E}[f(Y)\cdot X\mid Y=Y(\omega)]=\mathbb{E}[f(Y(\omega))\cdot X\mid Y=Y(\omega)]$$

And since $f(Y(\omega))$ is a constant, this is equal to:

$$=f(Y(\omega))\cdot\mathbb{E}[X\mid Y=Y(\omega)]=\Big(f(Y)\cdot\mathbb{E}[X\mid Y]\Big)(\omega)$$

As required.

■

Definition 2.9.8:

Given an event A and a random variable X with variance, we define **conditional variance** of X relative to A to be:

$$\mathrm{Var}(X\mid A)=\mathbb{E}\left[X^2\,\middle|\,A\right]-\mathbb{E}[X\mid A]^2$$

And if Y is another random variable, we define $\mathrm{Var}(X\mid Y)$ to be another random variable defined by:

$$\mathrm{Var}(X\mid Y)(\omega):=\mathrm{Var}(X\mid Y=Y(\omega))$$

Notice then that as a direct result of the definition

$$\mathrm{Var}(X\mid Y)=\mathbb{E}\left[X^2\,\middle|\,Y\right]-\mathbb{E}[X\mid Y]^2$$

Lemma 2.9.9:

$$\mathbb{E}[\mathrm{Var}(X\mid Y)]=\mathrm{Var}(X-\mathbb{E}[X\mid Y])$$

Proof:

Notice that

$$\mathbb{E}[\text{Var}(X \mid Y)] = \mathbb{E}\left[\mathbb{E}\left[X^2 \mid Y\right]\right] - \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right]$$

And

$$\text{Var}(X - \mathbb{E}[X \mid Y]) = \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X \cdot \mathbb{E}[X \mid Y]] + \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] - \left(\mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X \mid Y]]\right)^2$$

Notice that $\mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$, and that

$$\mathbb{E}[X \cdot \mathbb{E}[X \mid Y]] = \mathbb{E}[\mathbb{E}[X \cdot \mathbb{E}[X \mid Y] \mid Y]]$$

Now recall that **proposition 2.9.7** $\mathbb{E}[X \cdot \mathbb{E}[X \mid Y] \mid Y] = \mathbb{E}[X \mid Y] \cdot \mathbb{E}[X \mid Y]$ since $\mathbb{E}[X \mid Y]$ is a function of Y. So

$$\mathbb{E}[X \cdot \mathbb{E}[X \mid Y]] = \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right]$$

So all in all:

$$\text{Var}(X - \mathbb{E}[X \mid Y]) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right]$$

As required.

■

Theorem 2.9.10:

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathbb{E}[X \mid Y])$$

Proof:

Notice that

$$\mathbb{E}[\text{Var}(X \mid Y)] = \mathbb{E}\left[\mathbb{E}\left[X^2 \mid Y\right] - \mathbb{E}[X \mid Y]^2\right] = \mathbb{E}\left[\mathbb{E}\left[X^2 \mid Y\right]\right] - \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right]$$

And

$$\text{Var}(\mathbb{E}[X \mid Y]) = \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 = \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] - \mathbb{E}[X]^2$$

So all in all we get that

$$\mathbb{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathbb{E}[X \mid Y]) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \text{Var}(X)$$

As required.

■

3 Continuous Probability Spaces

3.1 General Probability Spaces

We will now generalize our discussion to not just focus on discrete probability spaces but general ones as well. This will be somewhat brief and we will then focus on continuous probability spaces. First let's revise our previous definition of a probability space.

Definition 3.1.1:

A **σ-algebra** is a set $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ for some set Ω which satisfies the following conditions:

(1) $\varnothing \in \mathcal{F}$

(2) If $\{U_n\}_{n\in\mathbb{N}} \in \mathcal{F}$ then their union is in $\mathcal{F}$ as well:

$$\bigcup_{n\in\mathbb{N}} U_n \in \mathcal{F}$$

(3) For every $U \in \mathcal{F}$, its complement is in $\mathcal{F}$ as well: $U^c \in \mathcal{F}$.

It then follows that $\Omega \in \mathcal{F}$ since it is the empty set's complement. $\mathcal{F}$ is also closed under finite unions since we can define $U_n = \varnothing$ for every n outside of the indexing set. And if $\{U_n\}_{n\in\mathbb{N}} \in \mathcal{F}$, then $U_n^c \in \mathcal{F}$ and therefore

$$\left(\bigcup_{n\in\mathbb{N}} U_n^c\right)^c \in \mathcal{F} \implies \bigcap_{n\in\mathbb{N}} U_n \in \mathcal{F}$$

So $\mathcal{F}$ is closed under intersections as well.

Proposition 3.1.2:

If I is an arbitrary indexing set and $\{\mathcal{F}_i\}_{i\in I}$ is a series of σ-algebras, then

$$\bigcap_{i\in I} \mathcal{F}_i$$

is also a σ-algebra.

Proof:

We will show that the intersection satisfies the conditions for being a σ-algebra.

(1) Since for every $\mathcal{F}_i$, $\varnothing \in \mathcal{F}_i$, $\varnothing$ is in the intersection as well.

(2) If $\{U_n\}_{n\in\mathbb{N}}$ is in the intersection, $\{U_n\}_{n\in\mathbb{N}} \in \mathcal{F}_i$ for every $i \in I$, so:

$$\bigcup_{n\in\mathbb{N}} U_n \in \mathcal{F}_i$$

Since $\mathcal{F}_i$ is a σ-algebra, and since this is true for every i, the union of U_n is in the intersection of $\mathcal{F}_i$.

(3) If U is in the intersection, it is in every $\mathcal{F}_i$, and therefore U^c is in every $\mathcal{F}_i$ as well and is therefore in the intersection.

■

So if we have a characteristic which we want a σ-algebra to have, then we can take the minimum σ-algebra which has this characteristic (minimum under $\subseteq$) by taking the intersection of all the σ-algebras which have this characteristic. The most famous and useful example of this are Borel Sets:

Definition 3.1.3:

Given a set $S \subseteq \mathbb{R}$, we define the **Borel Set** of S, $\mathbb{B}(S)$, to be the minimum σ-algebra which contains every closed

interval in S.
Furthermore, while this is redundant, we will require $\mathbb{B}(S) \subseteq \mathcal{P}(S)$.

$\mathbb{B}(S)$ exists since we can take the intersection of all the σ-algebras which contain every closed interval of S and are subsets of its powerset. This intersection is non-empty since $\mathcal{P}(S)$ is contained in it.

Definition 3.1.4:

A **Probability Space** is a triplet

$$(\Omega, \mathcal{F}, \mathbb{P})$$

Where Ω is a set called the **sample space**, $\mathcal{F}$ is a σ-algebra over Ω, and $\mathbb{P}$ is a probability function over $\mathcal{F}$, a function

$$\mathbb{P}\colon \mathcal{F} \longrightarrow [0,1]$$

Where $\mathbb{P}(\Omega) = 1$ and if $\{A_n\}_{n\in\mathbb{N}}$ are disjoint then

$$\mathbb{P}\left(\bigsqcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$$

Theorem 3.1.5:

There *exists* a probability function:

$$\mathbb{P}\colon \mathbb{B}([0,1]) \longrightarrow [0,1]$$

Where for every $[a,b] \subseteq [0,1]$, $\mathbb{P}([a,b]) = b - a$.

Notice that this doesn't tell us much about the actual definition of the Lebesgue measure, just a certain criteria it must fulfil. We will not be proving this theorem as we don't have the necessary tools to do so. To prove this we just need to show that the Lebesgue Measure satisfies the criteria.

Proposition 3.1.6:

Suppose $\mathbb{P}$ is the probability function defined over $\mathbb{B}([0,1])$ discussed above. Then

(1) $\mathbb{P}(\{a\}) = 0$

(2) $\mathbb{P}((a,b)) = b - a$

(3) If Q is countable, then $\mathbb{P}(Q) = 0$.

Proof:

(1) Notice that in **theorem 2.2.7**, we did not assume that the probability space was discrete, so we will use the result here. Let's define $A_n = \left[a, a + \frac{1}{n}\right]$ for every $n \in \mathbb{N}$. Then $\{A_n\}_{n\in\mathbb{N}}$ is decreasing, and its intersection is $\{a\}$. Furthermore, notice that $\mathbb{P}(A_n) = a + \frac{1}{n} - a = \frac{1}{n}$. So:

$$\mathbb{P}(\{a\}) = \mathbb{P}\left(\bigcap_{n\in\mathbb{N}} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \frac{1}{n} = 0$$

As required.

(2) Notice that:

$$\mathbb{P}([a,b]) = \mathbb{P}((a,b) \cup \{a\} \cup \{b\}) = \mathbb{P}((a,b)) + \mathbb{P}(\{a\}) + \mathbb{P}(\{b\}) = \mathbb{P}((a,b))$$

Since the probability of a singleton is 0. So:

$$\mathbb{P}((a,b)) = \mathbb{P}([a,b]) = b - a$$

(3) We know that

$$\mathbb{P}(Q) = \mathbb{P}\left(\bigsqcup_{q\in Q}\{q\}\right)$$

And since this is a disjoint countable union, this is equal to:

$$= \sum_{q\in Q}\mathbb{P}(\{q\}) = 0$$

■

Definition 3.1.7:

If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, a **random variable** is a function

$$X\colon \Omega \longrightarrow \mathbb{R}$$

Where for every $B \in \mathbb{B}(\mathbb{R})$, the preimage of B, $X^{-1}(B)$, is an event in $\mathcal{F}$. The reason for this is so $X \in B$ (which is the set $\left\{\omega \in \Omega \;\middle|\; \omega \in X^{-1}(B)\right\}$) is an event in $\mathcal{F}$.
The **distribution function** of X is a function

$$\mathbb{P}_X\colon \mathbb{B}(\mathbb{R}) \longrightarrow [0,1]$$

Such that $\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}\left(X^{-1}(B)\right)$. Note that $\mathbb{P}_X$ is a probability function over $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$.

Two random variables, X and Y, are **distributively equal** (or equivalent) if $\mathbb{P}_X = \mathbb{P}_Y$. This is denoted $X \overset{d}{=} Y$.

Definition 3.1.8:

If X is a random variable, then the **cumulative distribution** of X is the function

$$F_X\colon \mathbb{R} \longrightarrow [0,1]$$

Defined by

$$F_X(s) = \mathbb{P}(X \le s) = \mathbb{P}\left(X^{-1}\left((-\infty, s]\right)\right)$$

The **complementary cumulative distribution function** (or **tail distribution**) of X is the function

$$\bar{F}\colon \mathbb{R} \longrightarrow [0,1]$$

Defined by

$$\bar{F}(s) = \mathbb{P}(X > s) = 1 - F(s)$$

Proposition 3.1.9:

If X is a random variable, then:

(1) F_X is increasing.

(2) $\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$.

(3) F_X is continuous from the right and its limit exists from the left at every real point.

Proof:

(1) If $x < y$, then $\{X \le x\} \subseteq \{X \le y\}$, so

$$F_X(x) = \mathbb{P}(X \le x) \le \mathbb{P}(X \le y) = F_X(y)$$

As required.

(2) Let a_n be any monotonic increasing sequence to ∞. Then the sets $\{X \leq a_n\}$ are also increasing, and therefore

$$\lim_{n\to\infty} F_X(a_n) = \lim_{n\to\infty} \mathbb{P}(X \leq a_n) = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \{X \leq a_n\}\right)$$

And the union of all sets $\{X \leq a_n\}$ is Ω, since for every $\omega \in \Omega$, at some point $X(\omega) \leq a_n$. So:

$$= \mathbb{P}(\Omega) = 1$$

Since this is true for any monotonic increasing sequence to ∞, it must be true for any sequence whose limit is infinity, and therefore

$$\lim_{x\to\infty} F_X(x) = 1$$

Let a_n be any monotonic decreasing sequence to $-\infty$. Then the sets $\{X \leq a_n\}$ are also decreasing and therefore:

$$\lim_{n\to\infty} F_X(a_n) = \lim_{n\to\infty} \mathbb{P}(X \leq a_n) = \mathbb{P}\left(\bigcap_{n\in\mathbb{N}} \{X \leq a_n\}\right)$$

And the intersection of all sets $\{X \leq a_n\}$ is the empty set, as for every $\omega \in \Omega$, at some point $X(\omega) > a_n$. So

$$= \mathbb{P}(\varnothing) = 0$$

And for the same reason as above, this means

$$\lim_{x\to-\infty} F_X(x) = 0$$

(3) Let ε_n be a positive monotonic decreasing sequence to 0. Then:

$$\lim_{n\to\infty} F_X(x+\varepsilon_n) = \lim_{n\to\infty} \mathbb{P}(X \leq x+\varepsilon_n) = \mathbb{P}\left(\bigcap_{n\in\mathbb{N}} \{X \leq x+\varepsilon_n\}\right) = \mathbb{P}\left(\bigcap_{n\in\mathbb{N}} X^{-1}\Big((-\infty, x+\varepsilon_n)\Big)\right)$$

And this intersection is equal to $X^{-1}\Big((-\infty, x)\Big]$, so this is equal to:

$$= \mathbb{P}\Big(X^{-1}\Big((-\infty, x)\Big]\Big) = \mathbb{P}(X \leq x) = F_X(x)$$

And therfore for every $x_n \searrow x$, the limit of $F_X(x_n)$ is $F_X(x)$, so

$$\lim_{t\to x^+} F_X(t) = F_X(x)$$

And if ε_n is a negative increasing sequence to 0 then

$$\lim_{n\to\infty} F_X(x+\varepsilon_n) = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \{X \leq x+\varepsilon_n\}\right) = \mathbb{P}(X < x)$$

(Recall that by the definition of a limit, ε_n will never equal 0.) So

$$\lim_{t\to x^-} F_X(t) = \mathbb{P}(X < x)$$

So the limit exists.

■

Definition 3.1.10:

A random variable X is **discrete** if and only if there exists a countable set $B \in \mathbb{B}(\mathbb{R})$ such that $\mathbb{P}(X \in B) = 1$.

Note:

This is consistent with our previous definition, since singletons are in $\mathbb{B}(\mathbb{R})$, and B is countable:

$$\mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x) = 1$$

So we can define a mass probability function:

$$P_X : \Omega \longrightarrow [0,1]$$

By $P_X(x) = \mathbb{P}(X = x)$. Since for every $x \notin B$, $\mathbb{P}(X = x) = 0$ (otherwise the sum couldn't be 1), this creates a probability distribution.

And a random variable X is **continuous** if for every $x \in \mathbb{R}$, $\mathbb{P}(X = x) = 0$. Note that this means a random variable can't be both discrete and continuous.

Proposition 3.1.11:

A random variable is continuous if and only if its cumulative distribution function is continuous.

Proof:

On one hand, if X is continuous, then we need to show that F_X is left-continuous (since we already know it is right-continuous). We know that

$$\lim_{t \to x^-} F_X(t) = \mathbb{P}(X < x)$$

And we also know that $\mathbb{P}(X \leq x) = \mathbb{P}(X < x) + \mathbb{P}(X = x) = \mathbb{P}(X < x)$. Therefore F_X is also left continuous, as required. For the converse, we know that for every x:

$$\lim_{t \to x^-} F_X(t) = F_X(x) = \mathbb{P}(X \leq x)$$

Since F_X is continuous. But we know the limit is $\mathbb{P}(X < x)$, so $\mathbb{P}(X \leq x) = \mathbb{P}(X < x) + \mathbb{P}(X = x) = \mathbb{P}(X < x)$, so $\mathbb{P}(X = x) = 0$. Since this is true for every $x \in \mathbb{R}$, X is continuous.

■

Example:

It is possible for a random variable to be neither continuous nor discrete. Let X be a random variable over the probability space $([0,1], \mathbb{B}([0,1]), \mathbb{P})$ defined by $X(\omega) = \min\left\{\omega, \frac{1}{2}\right\}$. So $\mathbb{P}\left(X = \frac{1}{2}\right) = \mathbb{P}\left(\left[\frac{1}{2}, 1\right]\right) = \frac{1}{2}$, so X is not continuous. But suppose X is discrete, so there exists a countable S such that $\mathbb{P}(X \in S) = 1$. It is obvious then that $\frac{1}{2} \in S$. But for every other $x \neq \frac{1}{2}$, $\mathbb{P}(X = x) = 0$. So we get that:

$$1 = \sum_{x \in S} \mathbb{P}(X = x) = \frac{1}{2} + \sum_{\frac{1}{2} \neq x \in S} \mathbb{P}(X = x) = \frac{1}{2}$$

So we have that $1 = \frac{1}{2}$ in contradiction.

Lemma 3.1.12:

If X is a random variable and if $\{x_n\}_{n=1}^{\infty}$ is a sequence of distinct real points, then

$$\lim_{n \to \infty} \mathbb{P}(X = x_n) = 0$$

Proof:

We know that:

$$\mathbb{P}\left(X \in \bigcup_{n\in\mathbb{N}} x_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(X = x_n) \leq 1$$

And since $\mathbb{P}(X = x_n)$ is nonnegative, this sum must converge. And since the sum converges, the limit of $\mathbb{P}(X = x_n)$ must be 0, as required.

■

This means that for every $a \in \mathbb{R}$, the limit of $\mathbb{P}(X = x)$ as x approaches a is 0. This is because for every strictly monotonic sequence of points x_n to a, $\mathbb{P}(X = x_n)$ has a limit of 0. So the limit of $\mathbb{P}(X = x)$ is 0.
Notice then that X is continuous if and only if $\mathbb{P}(X = x)$ is continuous. This is true because if X is continuous then the limit of $\mathbb{P}(X = x)$ as x approaches a is 0, which is the same as $\mathbb{P}(X = a)$. And if $\mathbb{P}(X = x)$ is continuous then for every a, the limit of $\mathbb{P}(X = x)$ as x approaches a is $\mathbb{P}(X = a)$, and that limit is 0, so $\mathbb{P}(X = a) = 0$.

Proposition 3.1.13:

$$\mathbb{P}(X \in (a,b)) = \lim_{x\to a^+} \mathbb{P}(X \in (x,b)) = \lim_{x\to b^-} \mathbb{P}(X \in (a,x))$$

Proof:

We know that

$$\mathbb{P}(X \in (x,b)) = \mathbb{P}(X < b) - \mathbb{P}(X \leq x) = \mathbb{P}(X < b) - F_X(x)$$

And since F_X is right-conttinuous, as x approaches a from right, the limit of this is:

$$\lim_{x\to a^+} \mathbb{P}(X \in (x,b)) = \mathbb{P}(X < b) - F_X(a) = \mathbb{P}(X < b) - \mathbb{P}(X \leq a) = \mathbb{P}(X \in (a,b))$$

As required.
And for the other equality:

$$\mathbb{P}(X \in (a,x)) = \mathbb{P}(X \leq x) - \mathbb{P}(X \leq a) - \mathbb{P}(X = x) = F_X(x) - \mathbb{P}(X \leq a) - \mathbb{P}(X = x)$$

As x approaches b from the left, the limit of $F_X(x)$ is $\mathbb{P}(X < b)$, and by the lemma above the limit of $\mathbb{P}(X = x)$ is 0. So:

$$\lim_{x\to b^-} = \mathbb{P}(X < b) - \mathbb{P}(X \leq a) = \mathbb{P}(X \in (a,b))$$

As required.

■

Proposition 3.1.14:

$$\mathbb{P}(X \in (-\infty, a)) = \lim_{c\to-\infty} \mathbb{P}(X \in (c,a))$$

And

$$\mathbb{P}(X \in (a,\infty)) = \lim_{c\to\infty} \mathbb{P}(X \in (a,c))$$

Proof:

Recall that

$$\mathbb{P}(X \in (c,a)) = \mathbb{P}(X < a) - F_X(c)$$

And the limit of $F_X(c)$ as c goes to $-\infty$ is 0 as shown above, so the limit of this is $\mathbb{P}(X < a)$, which is $\mathbb{P}(X \in (-\infty, a))$ as required.

Notice that

$$\mathbb{P}(X \in (a,c)) = 1 - \mathbb{P}(X \leq a \text{ or } X \geq c) = 1 - \mathbb{P}(X \leq a) - \mathbb{P}(X \geq c) = \mathbb{P}(X < c) - \mathbb{P}(X \leq a) = F_X(c) - \mathbb{P}(X \leq a) - \mathbb{P}(X = c)$$

The limit of $F_X(c)$ is 1 as shown above, and the limit of $\mathbb{P}(X = c)$ is 0 as shown in the lemma above. So the limit of this is $1 - \mathbb{P}(X \leq a) = \mathbb{P}(X > a)$ which is $\mathbb{P}(X \in (a, \infty))$ as required.

■

Definition 3.1.15:

A random variable X is **absolutely continuous** if there exists a real nonnegative function f_X such that for every real $a < b$:

$$\mathbb{P}(X \in (a,b)) = \int_a^b f_X(x)\,dx$$

f_X is called X's **probability density function**.

Proposition 3.1.16:

If X is an absolutely continuous random variable then

(1) X is continuous.

(2)

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

(3)

$$F_X(s) = \int_{-\infty}^{s} f_X(x)\,dx \text{ and } \bar{F}_X(s) = \int_s^{\infty} f_X(x)\,dx$$

(4)

$$f(x) = F'(x)$$

Proof:

(1) Suppose $a \in \mathbb{R}$. Then:

$$\mathbb{P}(X = a) = \mathbb{P}\left(\bigcap_{n\in\mathbb{N}} \left\{X \in (a, a + \frac{1}{n})\right\}\right) = \lim_{n\to\infty} \mathbb{P}\left(X \in \left(a, a + \frac{1}{n}\right)\right) = \lim_{n\to\infty} \int_a^{a+\frac{1}{n}} f_X(x)\,dx$$

And this limit approaches 0. So $\mathbb{P}(X = a) = 0$ as required.

(2) Let x_n be a sequence of monotonically increasing values to infinity, and x_n' decreasing to negative infinity. Then:

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \lim \int_{x_n'}^{0} f_X(x) + \int_0^{x_n} f_X(x) = \lim \mathbb{P}(X \in (x_n', 0)) + \mathbb{P}(X \in (x_n, 0))$$

This is equal to

$$\lim \mathbb{P}(X \in (x_n', x_n))$$

Since X is continuous, so we can add in $\mathbb{P}(X = 0)$ as it is equal to 0. And this is an increasing series so this is equal to:

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \{X \in (x_n', x_n)\}\right) = \mathbb{P}(X \in \mathbb{R}) = 1$$

As required.

(3) We can use a similar proof as above for this.

(4) Suppose φ is an antiderivative of f_X (which must exist since f_X is integrable over every interval), then:

$$F_X(t) = \int_{-\infty}^{t} f_X(x)\,dx = \lim_{s\to-\infty} \varphi(t) - \varphi(s) = \varphi(t) - c$$

Where c is some constant. And if we differentiate both sides we get $F_X'(t) = f_X(t)$, as required.

■

Definition 3.1.17:

If if I is an interval with length ℓ (so $I = [a,b]$ or (a,b), etc. and $\ell = b-a$) and X is a random variable with a distribution:

$$f_X(x) = \frac{1}{\ell} \cdot \mathbb{1}_I(x)$$

Then X has a **uniform distribution** over I, this is denoted $X \sim \text{Unif}(I)$.

It is simple to verify that this is a valid probability density function, since its integral over $\mathbb{R}$ is equal to:

$$\int_a^b \frac{1}{b-a}\,dx = \frac{b-a}{b-a} = 1$$

Definition 3.1.18:

If X is a random variable with a probability density function:

$$f_X(t) = \lambda \cdot e^{-\lambda t} \cdot \mathbb{1}_{[0,\infty)}$$

Where $\lambda > 0$, then X has a **exponential probability distribution**, denoted $X \sim \text{Exp}(\lambda)$.

This is a probability density function since its integral over $\mathbb{R}$ is:

$$\int_0^\infty \lambda e^{-\lambda t}\,dt = e^{-\lambda t}\Big|_0^\infty = 1$$

3.2 Joint Probability, Expectation, and Variance

Definition 3.2.1:

If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a vector of absolutely continuous random variables then the **joint probability function** $f_{\boldsymbol{X}}$ is a function:

$$f_{\boldsymbol{X}} : \mathbb{R}^n \longrightarrow [0,1]$$

Where:

$$\mathbb{P}(\boldsymbol{X} \in (a_1, b_1) \times \cdots (a_n, b_n)) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f_{\boldsymbol{X}}(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

It can be shown with relative ease using **theorem 2.2.7** that if $X_1, \ldots, X_n$ have a joint probability function then:

$$\mathbb{P}(X_1 \leq a_1, \ldots, X_n \leq a_n) = \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, dx_n \cdots dx_1$$

Proposition 3.2.2:

If X and Y are two absolutely continuous random variables with a joint density function, X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ for every real x and y.

Proof:

If X and Y are independent then for every real a, b, c, and d:

$$\mathbb{P}(X \in (a,b), Y \in (c,d)) = \mathbb{P}(X \in (a,b)) \cdot \mathbb{P}(Y \in (c,d)) = \int_a^b f_X(x)\, dx \cdot \int_c^d f_Y(x)\, dx = \int_a^b \int_c^d f_X(x) \cdot f_Y(y)\, dy dx$$

So $f_X(x) \cdot f_Y(y)$ satisfies the property of the joint density function, so $f_{X,Y} = f_X \cdot f_Y$.
For the converse, we know that for every $I, J \in \mathbb{B}(\mathbb{R})$:

$$\mathbb{P}(X \in I, Y \in J) = \int_I \int_J f_{X,Y}(x,y)\, dy dx = \int_I \int_J f_X(x) \cdot f_Y(y)\, dy dx = \int_I f_X(x)\, dx \cdot \int_J f_Y(y)\, dy = \mathbb{P}(X \in I) \cdot \mathbb{P}(Y \in J)$$

So X and Y are independent, as required.

■

Proposition 3.2.3:

If X and Y are absoutely continuous random variables with a joint probability density function, for every real x:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$$

Proof:

Remember that:

$$\mathbb{P}(X \in (a,b)) = \mathbb{P}(X \in (a,b), Y \in \mathbb{R}) = \int_a^b \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy dx$$

And since this is true for every real a and b, if we define:

$$f_X(x) \coloneqq \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$$

We get that for every real $a < b$:

$$\mathbb{P}(X \in (a,b)) = \int_a^b f_X(x)\, dx$$

As required.

■

Theorem 3.2.4:

If X and Y are absolutely continuous random variables with a joint probability density function, and we define $Z = X + Y$, we have that

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(t, z-t)\,dt$$

is a probability density function of Z's.

Proof:

Notice that:

$$\mathbb{P}(X \leq a, Z \leq b) = \mathbb{P}(X \leq a, X + Y \leq b)$$

So we're integrating over when $X \leq a$ and $Y \leq b - X$:

$$= \int_{-\infty}^{a} \int_{-\infty}^{b-x} f_{X,Y}(x,y)\,dy\,dx$$

If we define $s = x + y$ then $dy = ds$ and $s(x, b-x) = b$ so:

$$= \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x, s-x)\,ds\,dx$$

If we then differentiate relative to a and then b we get that:

$$\frac{d}{db}\frac{d}{da}\mathbb{P}(X \leq a, Z \leq b) = \frac{d}{db}\int_{-\infty}^{b} f_{X,Y}(a, s-a)\,ds = f_{X,Y}(a, b-a)$$

But we know that this derivative is $f_{X,Z}(a,b)$, so:

$$f_{X,Z}(a,b) = f_{X,Y}(a, b-a)$$

And we know that:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Z}(x,z)\,dx = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\,dx$$

As required.

■

Definition 3.2.5:

If X and Y are two absolutely continuous random variables with a joint probability function, for every real y we define $X \mid Y = y$ to have a probability density function:

$$f_{X \mid Y=y} = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

If $f_Y(y) = 0$, we define this to just be 0.

Thus we get that:

$$\int_{-\infty}^{\infty} f_{X \mid Y=y}(x) \cdot f_Y(y)\,dy = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = f_X(x)$$

Which is the continuous version of **Law of Total Probability Version Two**.
Now we have arrived at the real purpose of this section, expectation. Once again we will define expected values, but for absolutely continuous random variables.

Definition 3.2.6:

If X is an absolutely continuous random variable with a density function f_X, then its expected value is defined to be:

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

If this integral converges absolutely.
And the variance of X is defined the same as before:

$$\mathrm{Var}(X) := \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2$$

Proposition 3.2.7:

If X has expectation then

$$\mathbb{E}[X] = \int_0^{\infty} \bar{F}_X(x)\, dx + \int_{-\infty}^{0} F_X(x)$$

Proof:

Notice that:

$$\int_0^{\infty} \bar{F}_X(x)\, dx = \int_0^{\infty} \int_x^{\infty} f_X(t)\, dt\, dx$$

This is integrating over $(t, x) \in \mathbb{R}^2$ where $0 \le x \le t$ so this is equal to the integral:

$$= \int_0^{\infty} \int_0^{t} f_X(x)\, dx\, dt = \int_0^{\infty} \left(\int_0^{t} dx \right) \cdot f_X(t)\, dt = \int_0^{\infty} t \cdot f_X(t)\, dt$$

Similarly we see that:

$$\int_{-\infty}^{0} F_X(x)\, dx = \int_{-\infty}^{0} t \cdot f_X(t)\, dt$$

And so we get that:

$$\int_0^{\infty} \bar{F}_X(x)\, dx + \int_{-\infty}^{0} F_X(x) = \int_{-\infty}^{\infty} t \cdot f_X(t)\, dt = \mathbb{E}[X]$$

■

Theorem 3.2.8 (The Law of the Unconscious Statistician):

If $\boldsymbol{X}$ is a vector of absolutely continuous random variables, and $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ is an integrable function such that $g^{-1}\big((a,b)\big) \in \mathbb{B}(\mathbb{R})^n$, then:

$$\mathbb{E}[g(\boldsymbol{X})] = \int \cdots \int_{\mathbb{R}^n} g(x_1, \ldots, x_n) \cdot f_{\boldsymbol{X}}(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

Specificaly we have that:

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

I will provide a proof for this specific case.

Proof:

Let us prove this for the simple case that g is a constant function over an interval $[a, b)$, that is $g(x) = \mathbb{1}_{[a,b)}(x)$. Then

$g(X) = \mathbb{1}_{[a,b]}(X)$, so $g(X)$ is 1 when $X \in [a,b)$ and 0 otherwise, this means that $g(X) \sim \text{Ber}(\mathbb{P}(X \in [a,b)))$. So:

$$\mathbb{E}[g(X)] = \mathbb{P}(X \in [a,b)) = \int_a^b f_X(x)\,dx = \int_{\mathbb{R}} g(x) \cdot f_X(x)\,dx$$

Since $g(x)$ is 0 for all reals that aren't in $[a,b)$.
Now suppose we have a countable partition $\left\{[x_{j-1}, x_j)\right\}_{j\in J}$ (that is $x_{j-1} < x_j$). We define $g_j = \mathbb{1}_{[x_{j-1},x_j)}$ for every $j \in J$. If then we have real numbers c_j and we define

$$g = \sum_{j\in J} c_j \cdot g_j$$

Then

$$g(X) = \sum_{j\in J} c_j \cdot g_j(X) = \sum_{j\in J} c_j \cdot \mathbb{1}_{[x_{j-1},x_j)}(X)$$

So $g(X)$ takes values c_j and

$$\mathbb{P}\Big(g(X) = c_j\Big) = \mathbb{P}\Big(X \in [x_{j-1}, x_j)\Big) = \int_{x_{j-1}}^{x_j} f_X(x)\,dx$$

So all in all we have that:

$$\mathbb{E}[g(X)] = \sum_{j\in J} c_j \cdot \int_{x_{j-1}}^{x_j} f_X(x)\,dx = \sum_{j\in J} \int_{x_{j-1}}^{x_j} c_j g_j(x) \cdot f_X(x)\,dx = \sum_{j\in J} \int_{x_{j-1}}^{x_j} g(x) \cdot f_X(x)\,dx = \int_{\mathbb{R}} g(x) \cdot f_X(x)\,dx$$

Now suppose that g is the sum of a countably infinite Now suppose we have a sequence of functions g^k which approach g from below which are constant over certain intervals, like above. So we can define:

$$g^k(x) = \inf_{t\in\left[n\cdot 2^{-k},(n+1)\cdot 2^{-k}\right]} g(t)$$

Where $n = \left\lfloor 2^k \cdot x \right\rfloor$, so for every t interval, $g^k(x)$ is constant. And this is a countable partition of $\mathbb{R}$ so by above:

$$\mathbb{E}\Big[g^k(X)\Big] = \int_{\mathbb{R}} g^k(x) \cdot f_X(x)\,dx$$

Also notice that since these partitions get finer, g^k is an increasing series, and it converges to g.
And since $g^k(X)$ is essentially a discrete random variable, $\mathbb{E}\Big[g^k(X)\Big] \leq \mathbb{E}\Big[g^{k+1}(X)\Big]$. By the monotone convergence theorem (which is a result of measure theory we will not be showing here), it turns out that:

$$\mathbb{E}\Big[g^k(X)\Big] \nearrow \mathbb{E}[g(X)]$$

And

$$\int_{\mathbb{R}} g_k(x) \cdot f_X(x)\,dx \nearrow \int_{\mathbb{R}} g(x) \cdot f_X(x)\,dx$$

And since we showed that the two series on the left are equal, since limits are unique, we get that:

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) \cdot f_X(x)\,dx$$

As required.

■

Theorem 3.2.9:

If X is a random variable with expectation (and variance when relevant) then

(1) If $X \overset{as}{\geq} 0$ then $\mathbb{E}[X] \geq 0$.

(2) $\mathbb{E}[\alpha X + \beta Y] = \alpha\,\mathbb{E}[X] + \beta\,\mathbb{E}[Y]$.

(3) If $X \overset{as}{\geq} Y$ then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

(4) If X and Y are independent $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

(5) $\mathrm{Var}(X) \geq 0$.

(6) $\mathrm{Var}(\alpha + X) = \mathrm{Var}(X)$ and $\mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X)$.

(7) If X and Y are independent then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof:

(1) So we have that the integral of $f_X(x)$ over $[0,\infty)$ is 1 and therefore $f_X(x) = 0$ almost always over $(-\infty, 0)$. Therefore:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx = \int_0^{\infty} x \cdot f_X(x)\,dx$$

Which is an integral of a nonnegative function and is therefore nonnegative.

(2) We know that by **The Law of the Unconscious Statistician**:

$$\mathbb{E}[\alpha X + \beta Y] = \iint_{\mathbb{R}^2} (\alpha x + \beta y) \cdot f_{X,Y}(x,y)\,dxdy = \alpha \iint_{\mathbb{R}^2} x \cdot f_{X,Y}(x,y)\,dxdy + \beta \iint_{\mathbb{R}^2} y \cdot f_{X,Y}(x,y)\,dxdy$$

Notice that:

$$\iint_{\mathbb{R}^2} x \cdot f_{X,Y}(x,y)\,dxdy = \int_{\mathbb{R}} x \cdot \int_{\mathbb{R}} f_{X,Y}(x,y)\,dydx = \int_{\mathbb{R}} x \cdot f_X(x)\,dx = \mathbb{E}[X]$$

And similar for y, so we get that this is equal to:

$$= \alpha\,\mathbb{E}[X] + \beta\,\mathbb{E}[Y]$$

As required.

(3) Since $X \overset{as}{\geq} Y$, $X - Y \overset{as}{\geq} 0$, so $\mathbb{E}[X-Y] = \mathbb{E}[X] - \mathbb{E}[Y] \geq 0$ and therefore $\mathbb{E}[X] \geq \mathbb{E}[Y]$ as required.

(4) Again by the law of the unconscious statistician:

$$\mathbb{E}[XY] = \iint_{\mathbb{R}^2} xy \cdot f_{X,Y}(x,y)\,dydx = \iint_{\mathbb{R}^2} xy \cdot f_X(x) \cdot f_Y(y)\,dydx = \int_{\mathbb{R}} x \cdot f_X(x) \cdot \int_{\mathbb{R}} y \cdot f_Y(y)\,dydx =$$

$$= \int_{\mathbb{R}} x \cdot f_X(x)\,dx \cdot \int_{\mathbb{R}} y \cdot f_Y(y)\,dy = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

(5) The proofs supplied in **2.7.2** are valid here since it assumes only the traits we proved above about expectation.

■

Also note that **Markov's Inequality**, **Chebyshev's Inequality**, **The Weak Law of Large Numbers**, and **theorem 2.7.7** also hold here since they only rely on these traits proved above.

Example:

If $X \sim \mathrm{Unif}[a,b]$ then:

$$F_X(t) = \int_a^t \frac{1}{b-a}\,dx = \frac{t-a}{b-a}$$

If $t \in [a,b]$ and if $t > b$ then it is 1 (since the integral is 1), and if $t < a$ it is 0. That is:

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

And its expected value is:

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

And notice that:

$$\mathbb{E}\left[X^2\right] = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{b^3-a^3}{3}$$

So:

$$\mathrm{Var}(X) = \frac{b^3-a^3}{3(b-a)} - \frac{(b-a)^2}{2} = \frac{(b-a)^2}{12}$$

Example:

If $X \sim \mathrm{Exp}(\lambda)$ then:

$$F_X(t) = \int_0^t \lambda \cdot e^{-\lambda x}\,dx = -e^{-\lambda x}\Big|_0^t = 1 - e^{-\lambda t}$$

For positive ts, and for negative ts it is 0, so:

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

Its expected value is:

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

And computing its variance gives $\mathrm{Var}(X) = \frac{1}{\lambda^2}$ (this is left as an exercise to the reader).

Theorem 3.2.10 (Memorylessness of Exponential Distributions):

If X is an absolutely continuous random variable, X distributes exponentially over λ if and only if $X - x_0 \mid X > x_0 \overset{d}{=} X$ for every nonnegative real x_0.

Proof:

In one direction, By definition we know that for every positive t:

$$\mathbb{P}(X - x_0 \geq t \mid X > x_0) = \frac{\mathbb{P}(X \geq t + x_0)}{\mathbb{P}(X > x_0)} = \frac{\bar{F}_X(t + x_0)}{\bar{F}_X(x_0)}$$

And we showed above that $\bar{F}_X(x) = e^{-\lambda x}$, so this is equal to:

$$= e^{-\lambda(t+x_0)+\lambda(x_0)} = e^{-\lambda t}$$

And so the cummulative probability distributions are equal and therefore $X - x_0 \mid X > x_0 \overset{d}{=} X$ as required.
In the other direction, notice that for every nonnegative real t:

$$\mathbb{P}(X \geq t) = \mathbb{P}(X - x_0 \geq t \mid X > x_0) = \frac{\mathbb{P}(X \geq t + x_0)}{\mathbb{P}(X > x_0)}$$

And since X is absolutely continuous, this means:

$$\bar{F}_X(t) \cdot \bar{F}_X(x_0) = \bar{F}_X(t + x_0)$$

Since this is true for every x_0, we can set $x_0 = t$ and we get that $\bar{F}_X(2t) = \bar{F}_X(t)^2$. Inductively, we can show that for every $n \in \mathbb{N}_1$, $\bar{F}_X(nt) = \bar{F}_X(t)^n$. Notice then that:

$$\bar{F}_X(t) = \bar{F}_X\left(n \cdot \frac{1}{n} \cdot t\right) = \bar{F}_X\left(\frac{1}{n} \cdot t\right)^n$$

So $\bar{F}_X(t)^{\frac{1}{n}} = \bar{F}_X\left(\frac{1}{n}\cdot t\right)$. Now suppose that $q \in \mathbb{Q}$ is a nonnegative rational, then there exists some naturals a and b such that $q = \frac{a}{b}$. So:

$$\bar{F}_X(q) = \bar{F}_X\left(\frac{a}{b}\right) = \bar{F}_X(a)^{\frac{1}{b}}$$

And since $\bar{F}_X(a) = \bar{F}_X(1)^a$, this is equal to $\bar{F}_X(1)^{\frac{a}{b}} = \bar{F}_X(1)^q$. So if we let $\bar{F}_X(1) = e^{-\lambda}$ for some real λ, we get that for every rational q: $\bar{F}_X(=)e^{-\lambda q}$. And since the rationals are dense in $\mathbb{R}$, it follows that this is true for every nonnegative real x.
Since this is a complementary cumulative probability distribution, this λ must be positive since its limit to infinity must be 0. And so this is the complementary cumulative distribution of an exponential distribution at least for nonnegative xs. But it must be equal to 1 for negative xs since it is decreasing and $\bar{F}_X(0) = 1$. So this is exactly the complementary cumulative probability distribution of an exponential distribution over λ, so $X - x_0 \mid X > x_0 \overset{d}{=} X$ as required.

■

This sheds light on an interesting connection between exponential and geometric distributions: they are both memoryless (by **Memorylessness of Geometric Distributions**). Another interesting connection is that if X has an exponential distribution, then $\lceil X \rceil$ has a geometric distribution! Let's prove this quickly.

Proof:

We know that $\mathbb{P}(\lceil X \rceil = x) = \mathbb{P}(X - 1 < X \leq x) = F_X(x) - F_X(x-1)$, and we can then apply the formula we found above for the cummulative distribution of exponential distributions above:

$$= 1 - e^{-\lambda x} - 1 + e^{-\lambda(x-1)} = e^{-\lambda(x-1)}\left(1 - e^{-\lambda}\right)$$

So if we define $p = 1 - e^{-\lambda}$ we get that:

$$\mathbb{P}(\lceil X \rceil = x) = (1-p)^{x-1} \cdot p$$

And therefore $\lceil X \rceil \sim \text{Geo}(p) = \text{Geo}\left(1 - e^{-\lambda}\right)$ as required.

■

Definition 3.2.11:

An absolutely continuous random variable X has a **normal distribution** over μ and σ^2 if it has a probability density function:

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

This is denoted $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

The normal distribution is one of the single most important distributions in all of probability. The reason why will become clear later on. But first lets investigate the distribution a bit.

Proposition 3.2.12:

If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ then $\alpha X + \beta \sim \mathcal{N}\left(\alpha\mu + \beta, \alpha^2 \cdot \sigma^2\right)$

Proof:

We know that:

$$\mathbb{P}(\alpha X + \beta \in (a,b)) = \mathbb{P}\left(X \in \left(\frac{a-\beta}{\alpha}, \frac{b-\beta}{\alpha}\right)\right) = \int_{\frac{a-\beta}{\alpha}}^{\frac{b-\beta}{\alpha}} \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sigma}\cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt$$

(Suppose the intervals here are bidirectional, ie $(a,b) = (b,a)$ so we don't have to worry about negative αs.) Let's substitute $u = \alpha t + \beta$. This means that $dt = \frac{du}{\alpha}$ so this is equal to:

$$= \int_a^b \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sigma\cdot\alpha}\cdot e^{-\frac{(u-(\alpha\mu+\beta))^2}{2\sigma^2\alpha^2}}\,du$$

So the probability density function of $\alpha X+\beta$ is exactly that of $\mathcal{N}\left(\alpha\mu+\beta,\alpha^2\cdot\sigma^2\right)$, as required. ■

So then if $Z\sim\mathcal{N}(0,1)$ (this is considered the "standard" normal distribution, or the normal normal distribution), and $X\sim\mathcal{N}\left(\mu,\sigma^2\right)$ then $X=\sigma\cdot Z+\mu$.

Proposition 3.2.13:

If $X\sim\mathcal{N}\left(\mu,\sigma^2\right)$ then $\mathbb{E}[X]=\mu$ and $\mathrm{Var}(X)=\sigma^2$.

Proof:

Let's focus on the specific case of $Z\sim\mathcal{N}(0,1)$. We know that:

$$\int_{-\infty}^{\infty}|t|\cdot e^{-\frac{t^2}{2}}\,dt\le\int_{-\infty}^{\infty}|t|\cdot e^{-|t|}\,dt$$

Which converges, as we know. So Z has an expected value.
Furthermore, we know that f_Z is symmetric about 0 ($f_Z(z)=f_Z(-z)$), so $t\cdot f_Z$ is odd, and therefore its integral over $\mathbb{R}$ is 0, so $\mathbb{E}[Z]=0$. And as we remarked above, $X\overset{d}{=}\sigma Z+\mu$, so $\mathbb{E}[X]=\sigma\,\mathbb{E}[Z]+\mu=\mu$, as required.
And Z^2 has expectation for the same reason as above, and by integration by parts:

$$\mathbb{E}\left[Z^2\right]=\frac{1}{\sqrt{2\pi}}\cdot\int_{\mathbb{R}}t^2e^{-\frac{t^2}{2}}\,dt=\frac{1}{\sqrt{2\pi}}\cdot\left(-te^{-\frac{t^2}{2}}\Big|_{-\infty}^{\infty}+\int_{\mathbb{R}}e^{-t^2}\,dt\right)$$

The rightmost integral is a famous integral called the Gaussian and has a known value of $\sqrt{2\pi}$, and the left term is equal to 0, so this is equal to:

$$=\frac{1}{\sqrt{2\pi}}\cdot\sqrt{2\pi}=1$$

And again $X\overset{d}{=}\sigma Z+\mu$ so $\mathrm{Var}(X)=\sigma^2\,\mathrm{Var}(Z)=\sigma^2$ as required. ■

Definition 3.2.14:

Due to its importance, the cumulative probability distribution of $Z\sim\mathcal{N}(0,1)$ gets a special symbol:

$$\Phi(t):=F_Z(t)$$

Notice that if $X\sim\mathcal{N}\left(\mu,\sigma^2\right)$ then:

$$F_X(t)=\mathbb{P}(X\le t)=\mathbb{P}(\sigma Z+\mu\le t)=\mathbb{P}\left(Z\le\frac{t-\mu}{\sigma}\right)=\Phi\left(\frac{t-\mu}{\sigma}\right)$$

3.3 Moment Generating Functions

Definition 3.3.1:

A real function f is **convex** in a set I if for every $a \in I$ there exists an m such that for every $x \in I$:

$$f(x) \geq f(a) + m(x-a)$$

Theorem 3.3.2:

If f is convex and X has an expected value, then:

$$f\Big(\mathbb{E}[X]\Big) \leq \mathbb{E}[f(X)]$$

Proof:

We know that there exists an m such that:

$$f(X) \geq f\Big(\mathbb{E}[X]\Big) + m\Big(X - \mathbb{E}[X]\Big)$$

Since $\mathbb{E}[X]$ is constant (it is our a). So if we take the expected value of both sides we get that

$$\mathbb{E}[f(X)] \geq \mathbb{E}\Big[f\Big(\mathbb{E}[X]\Big)\Big] + m\mathbb{E}[X - \mathbb{E}[X]] = f\Big(\mathbb{E}[X]\Big)$$

Since $f\Big(\mathbb{E}[X]\Big)$ is constant. As required.

■

Corollary 3.3.3:

If $\mathbb{E}\Big[X^k\Big]$ exists, then $\mathbb{E}\Big[X^{k-1}\Big]$ exists.

This means that if X has variance, it has expectation.

Proof:

Recall that $\mathbb{E}[Y]$ exists if and only if $\mathbb{E}[|Y|]$ exists. And it turns out that $x^{k/k-1}$ is convex (this is something you'd prove in calculus/analysis), so:

$$\mathbb{E}\Big[|X|^{k-1}\Big]^{\frac{k}{k-1}} \leq \mathbb{E}\Big[|X|^k\Big]$$

So if $\mathbb{E}\Big[X^k\Big]$ exists, then $\mathbb{E}\Big[|X|^k\Big]$ converges, and therefore so does $\mathbb{E}\Big[|X|^{k-1}\Big]$, as required.

■

Definition 3.3.4:

The kth **moment** of a random variable X is $\mathbb{E}\Big[X^k\Big]$, if it exists. And the **moment generating function** of X, denoted $M_X(t)$ is a function defined by:

$$M_X(t) = \mathbb{E}\Big[e^{t\cdot X}\Big]$$

For every t where this is defined.

Now it can be shown that expectation is linear even under infinite sums, but this requires a result from measure theory which we will not prove here. Therefore, we get that if every one of X's moments exists, then:

$$M_X(t) = \mathbb{E}\Big[e^{t\cdot X}\Big] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{t^k}{k!} \cdot X^k\right] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \cdot \mathbb{E}\Big[X^k\Big]$$

And therefore the nth derivative of X's moment generating function is:

$$M_X^{(n)}(t) = \sum_{k=n}^{\infty} \frac{k! \cdot t^{k-n}}{k! \cdot (k-n)!} \cdot \mathbb{E}\left[X^k\right] = \sum_{k=n}^{\infty} \frac{t^{k-n}}{(k-n)!} \cdot \mathbb{E}\left[X^k\right]$$

So if we let $t = 0$, the summands are all 0 except for when $k = n$, sp this becomes:

$$M_X^{(n)}(0) = \mathbb{E}[X^n]$$

So the moment generating function provides a powerful way of computing moments.

Proposition 3.3.5:

If X and Y are independent random variables, then $M_{X+Y} = M_X \cdot M_Y$.

Proof:

Notice that:

$$M_{X+Y}(t) = \mathbb{E}\left[e^{tX+tY}\right] = \mathbb{E}\left[e^{tX} \cdot e^{tY}\right]$$

And as we showed, if X and Y are inepended then so is $f(X)$ and $f(Y)$. So this is equal to:

$$= \mathbb{E}\left[e^{tX}\right] \cdot \mathbb{E}\left[e^{tY}\right] = M_X(t) \cdot M_Y(t)$$

As required. ■

Let's compute the moment generating functions of a few distributions.

- If $X \sim \text{Ber}(p)$ then notice that $e^{tX} = e^t$ with probability p and is 1 with probability $1-p$. So the moment generating function of X is
$$M_X(t) = p \cdot e^t + 1 - p$$
- If $X \sim \text{Bin}(n,p)$ then X is distributively equivalent to the sum of n independent bernoulli-distributing random variables with parameter p, and the moment generating function of a sum is the product of the moment generating functions, so:
$$M_X(t) = (p \cdot e^t + 1 - p)^n$$
- If $X \sim \text{Exp}(\lambda)$, then:
$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \int_0^{\infty} e^{tx} \cdot \lambda \cdot e^{-\lambda x}\, dx = \lambda \cdot \int_0^{\infty} e^{x(t-\lambda)}\, dx = \frac{\lambda}{t-\lambda} \cdot e^{x(t-\lambda)}\Big|_0^{\infty}$$
This only converges if $t < \lambda$ (if they're equal this just becomes the integral of λ which diverges). And if this is the case we get that:
$$M_X(t) = \frac{\lambda}{\lambda - t}$$
- If $X \sim \text{Geo}(p)$ then:
$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{k=1}^{\infty} e^{tk} \cdot p(1-p)^{k-1} = p \cdot e^t \cdot \sum_{k=1}^{\infty} \left(e^t(1-p)\right)^{k-1}$$
This converges if and only if $e^t(1-p) < 1$, that is $t < -\log(1-p)$. If this is the case then:
$$M_X(t) = \frac{pe^t}{1 - e^t(1-p)} = \frac{p}{e^{-t} + p - 1}$$

Theorem 3.3.6 (Chernoff Bound):

If X is a random variable, then for every positive real t where $M_X(t)$ defined, for every real a:

$$\mathbb{P}(X \geq a) \leq M_X(t) \cdot e^{-ta}$$

Proof:

By **Markov's Inequality**:

$$\mathbb{P}(X \geq a) = \mathbb{P}\left(e^{tX} \geq e^{ta}\right) \leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{ta}} = M_X(t) \cdot e^{-ta}$$

As required.

■

Lemma 3.3.7:

If X is a random variable such that $|X| \overset{as}{\leq} 1$ and $\mathbb{E}[X] = 0$ then for every real t:

$$M_X(t) \leq e^{\frac{t^2}{2}}$$

Proof:

The function $x \mapsto e^{tx}$ is convex (think second derivative), which means geometrically that the line between two points on the graph is above the function, so if we take the points $(1, e^t)$ and $(-1, e^{-t})$, we get that that for every $x \in [-1, 1]$:

$$e^{tx} \leq \frac{e^t - e^{-t}}{2} \cdot x + \frac{e^t + e^{-t}}{2}$$

So then since $X \in [-1, 1]$ almost surely:

$$M_X(t) = \mathbb{E}\left[e^{tx}\right] \leq \frac{e^t + e^{-t}}{2} \cdot \mathbb{E}[X] + \frac{e^t + e^{-t}}{2}$$

Since $\mathbb{E}[X] = 0$, this is equal to:

$$= \frac{e^t + e^{-t}}{2} \leq e^{\frac{t^2}{2}}$$

As required.

■

Theorem 3.3.8 (Hoeffding's Inequality):

If $\{X_k\}_{k=1}^n$ is a sequence of independent random variables such that for every k $\mathbb{E}[X_k] = 0$ and $|X_k| \overset{as}{\leq} 1$, then:

$$\mathbb{P}\left(\sum_{k=1}^{n} X_k \geq a\right) \leq e^{-\frac{a^2}{2n}}$$

Proof:

Let X be the sum of the X_ks. Then M_X is equal to the product of M_{X_k}s, so by the lemma above:

$$M_X(t) = \prod_{k=1}^{n} M_{X_k}(t) \leq \left(e^{\frac{t^2}{2}}\right)^n = e^{\frac{nt^2}{2}}$$

And by **Chernoff Bound**:

$$\mathbb{P}(X \geq a) \leq M_X(t) \cdot e^{-ta} \leq e^{\frac{nt^2}{2} - ta}$$

We define $f(t) = e^{\frac{nt^2}{2} - ta}$ and we will find its minimum, so we will find its derivative:

$$f'(t) = (nt - a)e^{\frac{nt^2}{2} - ta}$$

So $f'(t) = 0$ at $t = \frac{a}{n}$. So if we input that into the inequality above, we get:

$$\mathbb{P}(X \geq a) \leq e^{\frac{a^2}{2n} - \frac{a^2}{n}} = e^{-\frac{a^2}{2n}}$$

As required.

■

We can generalize this fact with the below corollary:

Corollary 3.3.9:

Suppose $\{X_k\}_{k=1}^n$ is a sequence of independent random variables such that there exists some M where for every k: $|X_k - \mathbb{E}[X_k]| \overset{as}{\leq} M$. Let X be the sum of the X_ks, then

$$\mathbb{P}(X - \mathbb{E}[X] \geq a) \leq e^{-\frac{a^2}{2nM^2}}$$

Proof:

Notice that:

$$\left|\frac{X_k - \mathbb{E}[X_k]}{M}\right| \leq 1$$

And

$$\mathbb{E}\left[\frac{X_k - \mathbb{E}[X_k]}{M}\right] = 0$$

So by the theorem above we get that:

$$\mathbb{P}(X - \mathbb{E}[X] \geq a) = \mathbb{P}\left(\frac{X - \mathbb{E}[X]}{M} \geq \frac{a}{M}\right) \leq e^{-\frac{a^2}{2nM^2}}$$

As required.

■