Probability Theory

Muchang Bahng

November 2022

An overview of probability using measures. We will denote probability measures defined over σ-algebras with $\mathbb{P}$ and probability functions defined over some sample space Ω or $\mathbb{R}$ with P or p.

1   Probability Spaces

Definition 1.1 (Probability Space). A **probability space** is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}(\Omega) = 1$.

1. Ω is called the **sample space** and an element $\omega \in \Omega$ is called an outcome.

2. $\mathcal{F}$ is called the **event space** and an element $A \in \mathcal{F}$ is called an event.

3. The measure of an event $\mathbb{P}(A)$ is called the **probability** of that event.

If some measure space X has a finite total measure, we can construct a probability space from it by normalizing the measure.

We can think of the sample space Ω as the set of all conceivable futures and an event $F \in \mathcal{F}$ as some subset of conceivable futures. The probability $\mathbb{P}(\Omega)$ represents our degree of certainty that our future will be contained in such an event. This formulation allows us to talk about discrete and continuous probability distributions at once. But given the same random experiment, we don't need to always have the same sample space. For example, let's have a coin toss. One could be interested in whether it lands heads or tails, which means $\Omega = \{0, 1\}$, but another could be interested in the number of times the coin flips midair, in which $\Omega = \mathbb{N}_0$. We could even be interested in the set of all trajectories of the coin, which would result in a huge space of all trajectories of the flip, or the velocity at which it lands on the table, which would lead to $\Omega = \mathbb{R}^+$.

Lemma 1.1 (Properties of Probability Measures). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then, the following properties hold:*

1. *Complement: If A is an event in $\mathcal{F}$,*

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

2. *Finite Additivity: If $A_1, \ldots, A_n$ are disjoint events, then*

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

3. *Monotonicity: If $A \subset B$, both in $\mathcal{F}$, then*

$$\mathbb{P}(A) \leq \mathbb{P}(B)$$

4. *Inclusion Exclusion Principle: If $A, B \in \mathcal{F}$,*

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

and by induction, if $A_1, \ldots, A_n \in \mathcal{F}$, then

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n-1} \mathbb{P}\left(\bigcap_{i=1}^n \mathbb{P}(A_i)\right)$$

5. *Continuity of probability: If $A_1, A_2, \ldots \in \mathcal{F}$, then*

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

Proof. Listed.

1. Since A and A^c are disjoint, it follows immediately from the axioms that $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$, and since they are all finite, we can get $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

2. Since $\mathcal{F}$ is by definition stable under countable union, use can fill a finite union up with empty sets, which are all disjoint, and then by definition the sum of their measures equal the measure of their unions.

3. We can use monotonicity to write

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

and since $\mathbb{P}$ is nonnegative by definition, $\mathbb{P}(B \setminus A) \geq 0$, proving the result.

4. Use set theory.

5. This is a highly nontrivial statement. Don't think of the infinite union $\cup_{i=1}^{\infty} A_i$ the sequential union of $A_1 \cup A_2 \cup \ldots$. Rather, just think of it as the set of all ω's that are in at least one of the A_i's. We can use the monotone convergence theorem to prove that the RHS is nondecreasing and bounded. So really we just have to prove equality. This is proved in Prop 2.5. of my measure theory notes, by defining $B_k = A_k - A_{k-1}$ and working with those.

$\blacksquare$

1.1   Discrete Probability Spaces

Definition 1.2 (Discrete Probability Space)**.** If Ω is a countable set, then we can take its σ-algebra $\mathcal{F}$ to be the power set of Ω and construct the measurable space $(\Omega, 2^{\Omega}, \mathbb{P})$. From the axioms, for any event $A \in \mathcal{F}$, we have

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) \text{ and } \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1$$

The greatest σ-algebra $F = 2^{\Omega}$ describes the complete information. The cases $\mathbb{P}(\{\omega\}) = 0$ is permitted by the definition, but rarely used since such ω can safely be excluded from the sample space. Therefore, we can define the probability measure $\mathbb{P}$ by simply defining it for all singleton sets $\{\omega\}$.

This may be confusing, since for discrete spaces, it looks like we're assigning probabilities to each $\omega \in \Omega$, but we are actually assigning them to singleton *sets*. We should be writing $\mathbb{P}(\{\omega\})$, but sometimes we abuse notation and write $\mathbb{P}(\omega)$.

Example 1.1. *Consider the flip of a fair coin with outcomes either hands or tails. Then, $\Omega = \{H, T\}$. The σ-algebra $F = 2^{\Omega}$ contains $2^2 = 4$ events:*

$$\{\} = \textit{Neither heads nor tails}$$
$$\{H\} = \textit{Heads}$$
$$\{T\} = \textit{Tails}$$
$$\{H, T\} = \textit{Either heads or tails}$$

That is, $\mathcal{F} = \{\{\}, \{H\}, \{T\}, \{H, T\}\}$. Our probability measure $\mathbb{P}$ is defined

$$\mathbb{P}(f) = \begin{cases} 0 & f = \{\} \\ 0.5 & f = \{H\} \\ 0.5 & f = \{T\} \\ 1 & f = \{H, T\} \end{cases}$$

Being able to consider the event space as 2^X is very nice, since countability of X allows us to avoid the Banach-Tarski paradox. It doesn't matter whether $\mathcal{F} = 2^X$ itself is uncountable or not.

Example 1.2. *A fair coin is tossed 3 times, creating 8 possible outcomes.*

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

The complete information is described by the σ-algebra $\mathcal{F} = 2^\Omega = 2^8 = 256$ events, where each of the events is a subset of Ω.

Example 1.3 (Geometric Measure on $\mathbb{N}$)**.** *Let $\Omega = \mathbb{N}$ and $\mathcal{F} = 2^{\mathbb{N}}$. We can completely define the probability measure by assigning them to singletons $k \in \mathbb{N}$. One such assignment is*

$$\mathbb{P}(\{k\}) = \frac{1}{2^k}$$

or more generally,

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

Example 1.4 (Poisson Measure on $\mathbb{N}_0$)**.** *Let $\Omega = \mathbb{N} \cup \{0\}$. Then, $\mathcal{F} = 2^\Omega$ and we can define $\mathbb{P}$ on the singleton sets as*

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

for any $\lambda > 0$. We can then compute the probability of, say all primes, by taking

$$\mathbb{P}(primes) = \sum_{k \ prime} \mathbb{P}(\{k\})$$

which we know to be monotonically increasing and bounded above, so it must converge. Whether this has a closed form solution is another matter. Again, in reality we are assigning probability measures on all $\mathcal{F}$-measurable sets, but just doing it through assignment of measure through singleton sets.

1.2   Continuous, Uncountable Sample Spaces

Now if we move to uncountable outcome spaces, then things are not as nice, which is why we need to machinery of measure theory to study them. Let us try to model a probability measure on $\Omega = [0, 1]$. It is uncountable, and it turns out that 2^Ω has cardinality strictly greater than even the continuum. If we try to model a uniform probability measure $\mathbb{P}$, then for some subset $A \in 2^\Omega$, it should be the case that $\mathbb{P}(A) = \mathbb{P}(A \oplus k)$, where $A \oplus k$ is just some translated version of A still contained within $[0, 1]$. This applies to singleton sets, and it turns out that if we try to assign a nonzero probability measure to any singleton $\{k\}$, then the probability measure of Ω blows up to infinity, which we can't have. So the only thing we can do is have every singleton have zero probability. Remember that a measure by definition has the *countable additivity* property, which says that

$$\mu\left(\bigsqcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k)$$

for all *countable* collections $\{A_k\}$. Summation is not defined for uncountable collections, and so having a probability 0 on every singleton does not imply that the probability of any uncountable set has is 0. That is, having $\mathbb{P}(\{k\}) = 0$ for all $k \in [0, 1]$ does not tell you what $\mathbb{P}([0, 1])$ is. So now rather than assigning probabilities to singletons, like we did with discrete sets, the approach is to assign probabilities directly to our event space $\mathcal{F}$. We can do this by directly assigning the Lebesgue measure to the Borel algebra of $[0, 1]$, which has the properties

1. $\mathbb{P}((a, b)) = \mathbb{P}([a, b)) = \mathbb{P}((a, b]) = \mathbb{P}([a, b]) = b - a$

2. Translation invariance as stated above.

Over uncountable Ω, we cannot afford to work with 2^Ω, since there is an impossibility theorem that says that there is no measure defined on $2^{[0,1]}$ with the two properties above. Therefore, we must work with a smaller σ-algebra. Since the subsets of interest are usually intervals (or more generally, open sets), people usually take the Borel σ-algebra of open intervals on $[0, 1]$. The Lebesgue measure on $\mathbb{R}$ is not a probability measure since it $\lambda(\mathbb{R}) = \infty$, but we can construct a uniform probability measure on any bounded set of $\mathbb{R}$. Usually, these continuous probability spaces are $\mathbb{R}^n$, and we define some measure μ directly on its σ-algebra.

1.2.1   Mixture Spaces

Definition 1.3 (Atom). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be uncountable. If for some $\omega \in \Omega$, $\mathbb{P}(\{\omega\}) \neq 0$, then ω is called an **atom**.

Now, given a general (discrete or continuous, or a combination of both) distribution, the set of all the atoms are an at most countable (maybe empty) set whose probability is the sum of probabilities of all atoms (by countable additivity). That is, given $\omega_1, \omega_2, \ldots$ atoms,

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

1. If this sum is equal to 1 then all other points can be safely excluded from the sample space Ω, returning us to the discrete case.

2. If this sum is 0 then we just have some continuous sample space. This means $\mathbb{P}(\{\omega\}) = 0$ for all $\omega \in \Omega$, and so Ω must be uncountable (since if it was countable, then we should be able to sum the $\mathbb{P}(\{\omega\})$'s to get 1, but it's 0). Remember that summation is only defined for at most countable elements.

3. If the sum of probabilities of all atoms is strictly between 0 and 1, then the probability space decomposes into a discrete, atomic part and a non-atomic, continuous part.

1.3   Infinite Coin Toss Model

Let us try to model a countably infinite sequence of coin tosses with $\Omega = \{0, 1\}^{\infty}$ and each $\omega \in \Omega$ of form

$$\omega = \omega_1 \omega_2 \omega_3 \ldots$$

which are infinite binary strings. Since Ω is uncountable, we can't just work with 2^{Ω}. We should try to assign an appropriate σ-algebra on Ω and then define a measure on it. Let's try to describe some sets that allows us to describe the outcome of the first n tosses. That is, $\mathcal{F}_n$ is a collection of subsets of Ω whose occurrence can be decided by looking at its first n tosses. Note that $\mathcal{F}_n$ is not a σ-algebra.

2   Conditional Probability

Definition 2.1 (Conditional Probability). Given a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, let B be an event such that $\mathbb{P}(B) > 0$. The **conditional probability** of A given B is defined

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

Note that we can't condition on events that have probability 0, which is why we need the $\mathbb{P}(B) > 0$ condition. If this is the case, it doesn't even make sense to talk about a conditional probability $\mathbb{P}(A \mid B)$. For example, if we take the probability space $[0, 1]$ with its Borel algebra and the Lebesgue measure, then we cannot condition something on the rationals, e.g. $\mathbb{P}(\{\omega < 0.5\} \mid \omega \in \mathbb{Q})$ does not make sense. In fact, doing so can lead to contradictions, one being the **Borel-Kolmogorov paradox**.

An extremely useful theorem is that the conditional probability taken as a measure gives us a new viable measure on the same probability space Ω.

Theorem 2.1. *Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Then, $\mathbb{P}(\cdot \mid B) : \mathcal{F} \longrightarrow [0, 1]$ is a probability measure on $(\Omega, \mathcal{F})$.*

Proof. We prove the properties of a probability measure.

1. The empty set has measure 0.

$$\mathbb{P}(\emptyset \mid B) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = \frac{0}{\mathbb{P}(B)} = 0$$

2. The entire space has measure 1.

$$\mathbb{P}(\Omega \mid B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$$

3. Countable additivity of disjoint events. Let $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$ which are disjoint. Then, their union is in $\mathcal{F}$ by definition of σ-algebra. Now,

$$
\begin{aligned}
\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i \Big| B \Big) &= \frac{1}{\mathbb{P}(B)} \mathbb{P}\Big[\Big(\bigcup_{i=1}^{\infty} A_i \Big) \cap B \Big] \\
&= \frac{1}{\mathbb{P}(B)} \mathbb{P}\Big[\bigcup_{i=1}^{\infty} (A_i \cap B) \Big] \\
&= \frac{1}{\mathbb{P}(B)} \sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B) \\
&= \sum_{i=1}^{\infty} \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i \mid B)
\end{aligned}
$$

∎

Theorem 2.2 (Partition Rule/Law of Total Probability). *Let $B_i \in \mathcal{F}$ for $i = 1, 2, \ldots$ with $\mathbb{P}(B_i) > 0$ be a partition of Ω and let $A \in \mathcal{F}$. Then,*

$$
\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i) \, \mathbb{P}(B_i)
$$

Proof. The left hand side is

$$
\mathbb{P}(A) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} (A \cap B_i) \Big) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i) \, \mathbb{P}(B_i)
$$

∎

Corollary 2.2.1. *If $B \in \mathcal{F}$ s.t. $0 < \mathbb{P}(B) < 1$, then*

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

Theorem 2.3 (Bayes Rule). *Let $A, B \in \mathcal{F}$. Then,*

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

Proof. We know that

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

and so we can write

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

∎

Corollary 2.3.1. *Let $A \in \mathcal{F}$ and $B_i \in \mathcal{F}$ for $i = 1, 2, \ldots$ be a partition of Ω. Then, using the partition rule,*

$$
\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i) \, \mathbb{P}(B_i)}{\sum_j \mathbb{P}(A \mid B_j) \, \mathbb{P}(B_j)}
$$

Theorem 2.4 (Probability of Intersection). *Let $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$. Then,*

$$
\mathbb{P}\Big(\bigcap_{i=1}^{\infty} A_i \Big) = \mathbb{P}(A_1) \prod_{i=2}^{\infty} \mathbb{P}\Big(A_i \,\Big|\, \bigcap_{j=1}^{i-1} A_j \Big)
$$

as long as the conditional probabilities are defined.

Proof. By continuity of probability (for first step), the left hand side equals

$$
\mathbb{P}\Big(\bigcap_{i=1}^{\infty} A_i \Big) = \lim_{n \to \infty} \mathbb{P}\Big(\bigcap_{i=1}^{n} A_i \Big) = \lim_{n \to \infty} \mathbb{P}(A_1) \prod_{i=2}^{\infty} \mathbb{P}\Big(A_i \,\Big|\, \bigcap_{j=1}^{i-1} A_j \Big)
$$

∎

3   Independence

Definition 3.1 (Independence of 2 Events). Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, events $A, B \in \mathcal{F}$ are said to be **independent under** $\mathbb{P}$ if
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$$
This leads to the immediate property that if $\mathbb{P}(B) > 0$, with A, B independent, then

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

Note that A and B may be independent under one measure, but not under another measure. The property that $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ is *not* the definition of independence, since it has the more restricting property that $\mathbb{P}(B) > 0$, so only refer to the definition that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. This is the true definition of independent events that we should rely on, not the one that says that A and B are independent if "one does not affect the other." This old definition is misleading and false. For example, take the probability space $[0, 1]$, with Borel σ-algebra, and Lebesgue measure $\mathbb{P} = \lambda$, and let $A = \mathbb{Q}$ and $B = \mathbb{R} \setminus \mathbb{Q}$. Then, contradictory to our old definition, A and B are independent since $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) = 0$! By the definition, an event A is independent of itself if $\mathbb{P}(A) = 0$ or 1 (e.g. A is rationals, irrationals, cantor set, $\emptyset$, Ω, etc.).

Definition 3.2 (Independence of n Events). Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$,

1. Let us have a finite collection of events $A_1, A_2, \ldots, A_n \in \mathcal{F}$. They are **independent** if for all nonempty $I_0 \subset \{1, 2, \ldots n\}$,
$$\mathbb{P}\left(\bigcap_{i \in I_0} A_i\right) = \prod_{i \in I_0} \mathbb{P}(A_i)$$

   Note that it is not enough to just prove that

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \prod_{i=1}^{n} \mathbb{P}(A_i)$$

   We must verify this for all 2^n possible choices (to be precise, we don't need to prove for $I_0 = \emptyset$ and $I_0 = \{A_i\}$), so for $2^n - n - 1$ choices.

2. Let $\{A_i\}_{i \in I}$ be a collection of events indexed by a possibly uncountable I. They are independent if for all nonempty and finite $I_0 \subset I$, we have

$$\mathbb{P}\left(\bigcap_{i \in I_0} A_i\right) = \prod_{i \in I_0} \mathbb{P}(A_i)$$

Now for a definition which is not seen in introductory probability courses.

Definition 3.3 (Sub-σ-Algebras). Given a σ-algebra $\mathcal{F}$, a **sub-σ-algebra** of $\mathcal{F}$ is a σ-algebra $\mathcal{G}$ s.t. $\mathcal{G} \subset \mathcal{F}$.

Now when we are trying to compare two σ-algebras, the measure defined for one may not even be defined on the other. To ensure that a measure is defined on both, it makes sense to take its σ-algebra and construct two sub-σ-algebras, which μ is guaranteed to be defined on.

Definition 3.4 (Independence of σ-Algebras). Let us have probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. Let $\mathcal{F}_1, \mathcal{F}_2$ be two sub-σ-algebras of $\mathcal{F}$. $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent if for any $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$, A_1 and A_2 are independent.

2. Let $\{\mathcal{F}_i\}_{i \in I}$ be an arbitrary collection of sub-σ-algebras of $\mathcal{F}$, indexed by possibly uncountable I. Then, they are independent if for any choices of $A_i \in \mathcal{F}_i$ for $i \in I$, $\{A_i\}_{i \in I}$ are independent events.

3.1 Borel-Cantelli Lemmas

There are many Borel-Cantelli lemmas, and we will introduce the two most famous ones. To understand what these lemmas say, given a sequence $A_1, A_2, \ldots$ of events in σ-algebra $\mathcal{F}$, we must first understand what the daunting term

$$\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i$$

means. Now let's try to explain what the intersection of the unions mean. First, remember that σ-algebras are stable under both countable unions and countable intersections, this is also in $\mathcal{F}$. We can interpret

$$\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i = \{A_n \text{ i.o.}\}$$

as the *event that infinitely many A_n's occur*, where i.o. means "infinitely often." To parse this, let's start from the innermost term and call it

$$B_n = \bigcup_{i=n}^{\infty} A_i \implies \{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty} B_n$$

B_n is the event that at least one of the $A_n, A_{n+1}, A_{n+2}, \ldots$ occurs, often referred to as the *nth tail event*. Now the intersection of all B_n's is the event that *all B_n's occur*. In other words, this is the event that for no matter how big of an $N \in \mathbb{N}$ I choose, there is always at least an event A_n with $n > N$ that occurs. This is shortly summarized as the event that infinitely many A_n's occur.

Lemma 3.1 (1st Borel-Cantelli Lemma). *Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, if $A_1, A_2, \ldots$ is a sequence of events such that*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$$

the almost surely (with probability 1) only finitely many A_n's will occur.

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i \right) = 0$$

Proof. Setting B_n as above, we have

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n \right) = \lim_{n \to \infty} \mathbb{P}(B_n) \qquad \text{(continuity of probability)}$$

$$= \lim_{n \to \infty} \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \right) \qquad \text{(substitute } B_i\text{)}$$

$$\leq \lim_{n \to \infty} \sum_{i=n}^{\infty} \mathbb{P}(A_i) = 0 \qquad \text{(tail sum of convergent series is 0)}$$

$\blacksquare$

The second Borel-Cantelli lemma is like a partial contrapositive to the first lemma, where it starts with the assumption that the sum of the $\mathbb{P}(A_n)$'s are infinite (along with the addition case that they are independent).

Lemma 3.2 (2nd Borel-Cantelli Lemma). *If $A_1, A_2, \ldots$ are independent events such that*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty,$$

then almost surely (with probability 1) infinitely many A_n's will occur. That is,

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

The intuition behind this lemma is challenging: We can let $\mathbb{P}(A_n) = P_n$ and interpret the sum as a series of P_n's. Since the series $P_1 + P_2 + \ldots$ is finite, this implies that

$$\lim_{n \to \infty} P_n = 0$$

(but not the converse) and going to zero rather fast such that the series is finite. So, you are working with a sequence of events A_n that are becoming more and more unlikely rather fast. The lemma says that beyond a certain point n_0, none of the events A_n will occur almost surely. For the second lemma, we can go as far as we like in the sequence of A_n's, up to any A_{n_0}, but beyond that there is always an infinite number of A_n's that occur beyond A_{n_0}.

4   Random Variables

Random variables are motivated by the following. When you have a random experiment, the experimenter may not be interested in the specific elementary outcomes. So if you have sample space Ω, you may not be concerned about what $\omega \in \Omega$ shows up, but more interested in some numerical function of the elementary outcome. For example, if you toss a coin 10 times, you're not interested in what sequence in $\{0, 1\}^{10}$ shows up, but you may want to just know how many heads came up. In other words, your interest defines a numerical function $X : \Omega \to \mathbb{R}$. This is useful, since in many cases the sample space Ω can be extremely complicated (e.g. the sample space of all weather conditions) and the elementary outcomes also complicated, so you may want to know some simpler aspect (e.g. the temperature).

The name "random variable" is very misleading. It's not random nor a variable. It is a deterministic function $X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow \mathbb{R}$ that assigns numbers to outcomes. The only source of randomness itself is which $\omega \in \Omega$ is chosen. But we can't just choose any function on Ω; they must satisfy the nice property of measurability. Now, to talk about random variables, recall that the definition of a measurable function $f : (X, \mathcal{A}) \longrightarrow \mathbb{R}$ is one where the preimage of every Borel set $B \in \mathcal{B}(\mathbb{R})$ is in $\mathcal{A}$. With a potential measure μ, this allows us to define the Lebesgue integral of f. Note that this is also equivalent to the more easily provable fact that the preimage of every half-interval $(-\infty, t)$ is in $\mathcal{A}$. That is, $f^{-1}((-\infty, t]) \in \mathcal{A}$ for all $t \in \mathbb{R}$.

Definition 4.1 (Random Variable). A **random variable** X on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is an $\mathcal{F}$-measurable function $X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow \mathbb{R}$. That is, for every subset $B \in \mathcal{B}(\mathbb{R})$, its preimage

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}$$

The reason we want X to be $\mathcal{F}$-measurable is because now we can define probabilities on Borel sets B of $\mathbb{R}$ by computing the probabilities of the preimage of B, which must be $\mathcal{F}$-measurable. In a way, a random variable "pushes forward" the probability measure $\mathbb{P}$, originally defined on $\mathcal{F}$, to $\mathcal{B}(\mathbb{R})$.

Definition 4.2 (Probability Law of Random Variable X). Let X be a random variable on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The **probability law of** X is a function $\mathbb{P}_X : \mathcal{B}(\mathbb{R}) \longrightarrow [0, 1]$ defined, for each Borel set B of $\mathbb{R}$, as

$$\mathbb{P}_X(B) := \mathbb{P}\big(X^{-1}(B)\big) = \mathbb{P}\big(\{\omega \in \Omega \mid X(\omega) \in B\}\big)$$

Note that $\mathbb{P}$ refers to the probability measure on $\mathcal{F}$, and $\mathbb{P}_X$ refers to the probability law on $\mathcal{B}(\mathbb{R})$. In shorthand, we can write $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$. By abuse of notation, it is generally written $\mathbb{P}(X \in B)$.

Note that if we are working in a discrete probability space Ω, then we can simply take the σ-algebra to be 2^{Ω}, and so we can take any function on Ω as a random variable since its preimage will always be in 2^{Ω}.

Example 4.1 (Coin Flip). *We start off with a coin flip. Let Ω denote the set of all conceivable futures of this coin flip: all ways it can fly, spin, or bounce before coming to rest. Then, we can define a random variable X that assigns 0 to all outcomes landing on tails and 1 to all outcomes landing on heads. For example, there could be two trajectories $\omega_1, \omega_2 \in \Omega$ that the coin could take before landing on heads. Then, $X(\omega_1) = X(\omega_2) = 1$. Then, $\{1\} \in \mathcal{B}(\mathbb{R})$, and*

$$\mathbb{P}[X = 1] = \mathbb{P}[X^{-1}(\{1\})] = \mathbb{P}[\{\omega \in \Omega \mid X(\omega) = 1\}] = \frac{1}{2}$$

Theorem 4.1. *Given* $(\Omega, \mathcal{F}, \mathbb{P})$ *and a random variable* $X : \Omega \to \mathbb{R}$, *let us define a probability law* $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$. *Then,*

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$$

is a probability space.

This theorem is extremely useful, since in practical applications, one does not consider an abstract Ω and works immediately in $\mathbb{R}$. Once we have determined our numerical values of interest (heads or tails, number of heads, sum of dice rolls) with our random variable X, we can just throw away $(\Omega, \mathcal{F}, \mathbb{P})$ and work directly in probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$.

4.1   Cumulative Distribution Function

Now, remember that the Borel algebra $\mathcal{B}(\mathbb{R})$ is generated by the semi-infinite intervals of form $(-\infty, t]$ (for all $t \in \mathbb{R}$), which are considered "nice" Borel sets. So, $\mathbb{P}_X((-\infty, t])$ is well defined for all $t \in \mathbb{R}$. In fact, this has a name.

Definition 4.3 (Cumulative Distribution Function). Given $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$. Then, the **cumulative distribution function** of X is defined

$$F_X(x) = \mathbb{P}\big(\{\omega \in \Omega \mid X(\omega) \leq x\}\big)$$

We can also define this with the probability law $\mathbb{P}_X$ as

$$F_X(x) = \mathbb{P}_X\big((-\infty, x]\big)$$

] By abuse of notation, we will write the CDF as $P(X \leq x)$. It satisfies the properties:

1. Limits:
$$\lim_{x \to -\infty} F_X(x) = 0 \text{ and } \lim_{x \to \infty} F_X(x) = 1$$

2. Monotonicity:
$$x \leq y \implies F_X(x) \leq F_X(y)$$

3. Right-continuity: For all $x \in \mathbb{R}$
$$\lim_{\epsilon \to 0^+} F_X(x + \epsilon) = F_X(x)$$

So, if there are jumps, the hole can exist as the function approaches a value from the left.

What is remarkable is that any function satisfying these three properties satisfies these 3 properties gives you a viable CDF (and as shown below, completely determines a unique random variable).

So if you give me the probability law for all Borel sets of $\mathbb{R}$, then I can easily define the CDF since $(-\infty, x]$ are also Borel sets. It turns out that if we know *just* the CDF, then since the semi-infinite intervals form a generating class of $\mathcal{B}(\mathbb{R})$, it turns out that we can completely define $\mathbb{P}_X$. The proof of the theorem below is a bit more involved, using π-systems, but it is good to know.

Theorem 4.2. *The CDF* $F_X(\cdot)$ *uniquely specifies the probability law* $\mathbb{P}_X$ *for any random variable* X.

To summarize, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable just pushes a measure onto the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If we only care about the values of the random variable, then we can forget about Ω and only look at $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$. The CDF on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ will be well defined since semi-finite intervals are also Borel. If I am just given a CDF $F_X(\cdot)$, then this is enough for me to specify a unique probability measure $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. So although $\mathbb{P}_X$ contains the complete description of the random variable X, in practice we will use F_X since it also captures all the information of X and it's easier to work with.

Definition 4.4 (Indicator Random Variable). Given $(\Omega, \mathcal{F}, \mathbb{P})$, let $A \in \mathcal{F}$ be an event. A useful random variable is the **indicator random variable** $I_A : \Omega \longrightarrow \mathbb{R}$ defined

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

This is a random variable since the preimages of $\emptyset, \{0\}.\{1\}, \{0,1\}$ are $\emptyset, A^c, A, \Omega$, which are all $\mathcal{F}$-measurable. The CDF of this function will look like a step function

$$F_{I_A}(x) = \begin{cases} 0 & \text{if } x < 0 \\ P(A^c) & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x \end{cases}$$

4.2　Discrete Random Variables

You classify random variables based on the nature of the measure $\mathbb{P}_X$ induced on the real line. There are only three fundamental types of measures: discrete, continuous and singular random variables. In fact, a result in measure theory called *Lebesgue's Decomposition Theorem* says that every measure on $\mathbb{R}$ are either one of these 3 or mixtures thereof. We are used to the first two; the third one is very bizzare and has little to no practical applications.

Definition 4.5 (Discrete Random Variable). Given $(\Omega, \mathcal{F}, \mathbb{P})$, let us have a random variable X that induces a probability law on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. X is said to be **discrete** if there exists a countable set $E \subset \mathbb{R}$ s.t. $\mathbb{P}_X(E) = 1$ (i.e. E's preimage has probability measure 1). Since E is at most countable, we can enumerate it $E = \{e_1, e_2, \ldots\}$, and by countable additivity of disjoint sets, we have

$$1 = \mathbb{P}_X(E) = P\Big(\bigcup_{i=1}^{\infty} \{e_i\} \Big) = \sum_{i=1}^{\infty} \mathbb{P}_X(\{e_i\}) = \sum_{i=1}^{\infty} P(X = e_i)$$

and for any $B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_X(B) = \sum_{x \in E \cap B} P(X = x)$$

Therefore, the entire probability measure is determined by the probabilities of the singleton sets $P(X = e_i)$. Therefore, the function

$$p_X(x) := P(X = x)$$

is called the **probability mass function** of X, and we can compute using the Lebesgue integral, which reduces to the summation:

$$\mathbb{P}_X(B) = \int_B p_X(x) \, d\mathbb{P}_X = \sum_{x \in E \cap B} p_X(x)$$

Sometimes, the definition of discrete X involves having a countable image in $\mathbb{R}$, but our definition allows us to have some $B \in \mathcal{B}(\mathbb{R})$ where its preimage is not necessarily the sample space Ω, but a smaller subset of measure 1. What's nice about the discrete random variable is that the probability mass function p_X completely describes its probability law. The CDF of a discrete probability function will look like an increasing series of steps. If we have $E = \{e_1, e_2, e_3, e_4, e_5\}$, its CDF would look like:
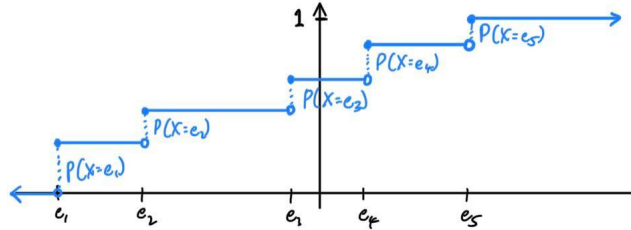
If E was countable, then it would have countably infinite discontinuities. Now we'll give some examples of discrete random variables, and in here we'll completely ignore the sample space Ω, since once we have a random variable X, we can just work in $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$. Remember that we will write $P(X = x)$ as shorthand for $\mathbb{P}_X(\{x\})$.

Example 4.2 (Bernoulli Random Variable). *Given whatever sample space, a Bernoulli random variable $X : \Omega \longrightarrow \mathbb{R}$, denoted $X \sim \mathrm{Bernoulli}(p)$, has $E = \{0, 1\}$ and a PMF defined*

$$p_X(x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

This PMF completely defines the probability law $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$, and we can see that $\mathbb{P}_X(E) = \mathbb{P}_X(\{0\}) + \mathbb{P}_X(\{1\}) = P(X = 0) + P(X = 1) = 1$.

Example 4.3 (Uniform Random Variable). *Given a finite set $E = \{e_i\}_{i=1}^n \subset \mathbb{R}$, we define the PMF as*

$$p_X(e_i) = \mathbb{P}(X = e_i) = \frac{1}{n} \; \forall i = 1, 2, \ldots n$$

which induces the probability measure $\mathbb{P}_X(B) = \sum_{x \in E \cap B} p_X(x)$.

The Bernoulli RV leads to the geometric and binomial random variables.

Example 4.4 (Geometric Random Variable). *Given $E = \mathbb{N}$, we can define the PMF associated with random varibale $X \sim \text{Geometric}(p)$ as*

$$p_X(k) = \mathbb{P}(X = k) = (1-p)^{k-1}p \text{ for } k \in \mathbb{N}, \ p \in [0,1]$$

which induces the probability measure $\mathbb{P}_X(B) = \sum_{x \in E \cap B} p_X(x)$. We can interpret this as the number of times you have to (independently) toss a p-coin (probability of heads is p) until you get a heads.

Example 4.5 (Binomial Random Variable). *We let $E = \mathbb{N}_0$ and define the PMF associated with random variable $X \sim \text{Binomial}(n,p)$ as*

$$p_X(k) = \mathbb{P}(X = k) = \binom{n}{k}p^k(1-p)^{n-k} \text{ for } k \in E, p \in [0,1]$$

We can interpret this as the number of heads occurring in a sequence of n independent tosses of a p-coin.

Example 4.6 (Poisson Random Variable). *We let $E = \mathbb{N}_0$ and define the PMF of $X \sim \text{Poisson}(\lambda)$ as*

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k \in E, \ \lambda > 0$$

4.3 Simple Random Variables

A slight generalization of a discrete random variable is a simple random variable. Recall that the indicator random variable is a function $\mathbb{I}_A : \Omega \to \mathbb{R}$ defined

$$\mathbb{I}_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if else} \end{cases}$$

As simple random variable generalizes this into multiple sets that form a partition of Ω. It is analogous to a simple function, introduced in measure theory.

Definition 4.6 (Simple Random Variable). Let $\{A_i\}_i$ form a partition of probability space Ω. A **simple random variable** X is a random variable of the form

$$X(\omega) = \sum_i a_i \mathbb{I}_{A_i}(\omega)$$

that assigns value a_i if the input $\omega \in A_i$.

4.4 Continuous Random Variables

Definition 4.7 (Absolutely Continuous Measures). Let μ, ν be measures defined on $(\Omega, \mathcal{F})$. We say that ν is **absolutely continuous** w.r.t. μ if for every $N \in \mathcal{F}$ s.t. $\mu(N) = 0$, we have $\nu(N) = 0$.

Definition 4.8 (Continuous Random Variable). A random variable X is **continuous** if its induced measure $\mathbb{P}_X : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to [0,1]$ is absolutely continuous w.r.t. the Lebesgue measure $\lambda : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to \mathbb{R}$, i.e. if for every Borel set N of Lebesgue measure 0, we have $\mathbb{P}_X(N) = 0$ also.

A common misconception is that a random variable X is continuous if the induced measure on every singleton set in $\mathcal{B}(R)$ is 0, i.e. $\mathbb{P}_X(\{x\}) = 0$ for all $x \in \mathbb{R}$. The definition above implies this since the Lebesgue measure of every singleton set is 0.

We introduce a theorem that is useful to know, but we won't prove it.

Theorem 4.3 (Radon-Nikodym Theorem (Special Case)). *Let X be a continuous random variable. Then, there exists a nonnegative measurable function $f_X : \mathbb{R} \longrightarrow [0, \infty)$ s.t. for any $B \in \mathcal{B}(\mathbb{R})$, we have*

$$\mathbb{P}_X(B) = \int_B f_X \, d\lambda$$

*where the above is the Lebesgue integral. Note that we must define using the Lebesgue integral because Riemann integral is not compatible with any Borel set. f_X is called the **probability density function**,*

aka **PDF**. *Furthermore, we can get* f_X *from* $\mathbb{P}_X$ *by taking the* **Radon-Nikodym derivative** *(which we will not define now)*

$$f_X = \frac{d\mathbb{P}_X}{d\lambda}$$

which basically says that if we have a set of very small Lebesgue measure $d\lambda$ *tending to* 0*, then its probability measure* $\mathbb{P}_X$ *will also be very small, and the infinitesimal ratio of these two measures on an arbitrarily small set is* f_X*. Also, note that the integral does not change if the value of* f *changes on sets of Lebesgue measure* 0*, and so there is no unique pdf describing* $\mathbb{P}_X$*. It is unique up to sets of Lebesgue measure* 0*, so when we refer to such a pdf* f_X*, we are really talking about an equivalence class of functions.*

This theorem guarantees the existence of some f_X that completely describes the probability law P_X! Take a special case of when $B = (-\infty, x])$, and we can define the CDF as

$$F_X(x) = P_X((-\infty, x]) = \int_{(-\infty, x]} f_X \, d\lambda$$

If the set of integration is an interval (and the function is continuous a.e.), then the Lebesgue integral and Riemann integral coincides, and we get the familiar formula

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt$$

and we can differentiate it to get back the pdf f_X (or more accurately, some function that agrees with f_X a.e.). We can show that the CDF of a continuous random variable X

1. is absolutely continuous, and

2. is differentiable almost everywhere, which means that its pdf will be defined almost everywhere (and we can fill in the undefined points however we want).

Note that the pdf f_X itself has no interpretation as a probability (indeed, we can change its value at a countable number of points to anything we want). It is only when we integrate it over some Borel set that gives us a probability.

Example 4.7 (Uniform Random Variable)**.** *Let us define the uniform probability measure* P_X *on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *with the CDF*

$$F_X = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } 1 < x \end{cases}$$

It is differentiable almost everywhere except for at the two points $x = 0$ *and* $x = 1$*. Therefore, the pdf* f_X *is defined for all real numbers except* $x = 0$ *and* $x = 1$*. But it doesn't matter: we can assign any value* f_X *we want on* 0 *and* 1 *since it won't affect the integral of it. In this example, we just set*

$$f_X = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if else} \end{cases}$$

Example 4.8 (Exponential Random Variable)**.** *The exponential random variable has the following cdf:*

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

which is differentiable everywhere except at $x = 0$*. Differentiating it (and assigning a convenient value at* $x = 0$ $f(0) = \lambda$*) gives the pdf*

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

The exponential random variable has the memoryless property.

Definition 4.9 (Memoryless Property). A random variable X has the **memoryless property** if it satisfies for all $t, s \geq 0$

$$\mathbb{P}(X > s + t \mid x > t) = \mathbb{P}(X > s)$$

which is just abuse of notation for the following: We know that (t, ∞), (s, ∞), and $(s + t, \infty)$ are all in $\mathcal{B}(\mathbb{R})$ and so they are events. So it really translates to the probability of an outcome landing in $(s+t, \infty)$ given that it lands in (t, ∞) is equal the probability of it landing in (s, ∞).

$$\mathbb{P}_X\big((s + t, \infty) \mid (t, \infty)\big) = \frac{\mathbb{P}_X\big((s + t, \infty) \cap (t, \infty)\big)}{\mathbb{P}_X\big((t, \infty)\big)} = \frac{\mathbb{P}_X\big((s + t, \infty)\big)}{\mathbb{P}_X\big((t, \infty)\big)} = \mathbb{P}_X\big((s, \infty)\big)$$

The exponential random variable is memoryless because the LHS just reduces to

$$\frac{\mathbb{P}_X\big((s + t, \infty)\big)}{\mathbb{P}_X\big((t, \infty)\big)} = \frac{1 - F_X(s + t)}{1 - F_X(t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = 1 - F_X(s) = \mathbb{P}_X\big((s, \infty)\big)$$

In fact, the only continuous random variable having the memoryless property is the exponential random variable.

Example 4.9 (Gaussian Random Variable). *The pdf is easier to specify for the Gaussian, so we define the Gaussian RV as having pdf*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \text{ for } \mu \in \mathbb{R}, \sigma > 0$$

*Note that this pdf decreases very quickly as we get further from μ. The cdf cannot be written in closed form, and we call the cdf of the standard Gaussian the **error function**:*

$$\mathrm{Erf}(x) = F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt$$

Example 4.10 (Cauchy Random Variable (Standardized)). *The Cauchy random variable gives the pdf*

$$f_X(x) = \frac{1}{\pi} \frac{1}{1 + x^2} \text{ for } x \in \mathbb{R}$$

Integrating this gives the inverse tangent, which after scaling it down by π satisfies the conditions of the cdf. Note that the Cauchy distribution falls off much more slowly around the mean (at a rate of $\frac{1}{1+x^2}$, like a power law) than the Gaussian (which is even faster than an exponential, it is at the rate of e^{-x^2}). If such a pdf falls off at a slow rate, like a power law, then this is called a heavy-tailed random variable.

Example 4.11 (Gamma Random Variable). *The pdf associated with random variable $X \sim \mathrm{Gamma}(n, \lambda)$ is defined*

$$f_X(x) = \frac{\lambda^n x^{n-1}}{\Gamma(n)} e^{-\lambda x} \text{ for } x \geq 0$$

where Γ is the gamma function, which is an extension of the factorial function to the domain of complex numbers.

$$\Gamma(x) := \int_0^{\infty} z^{x-1} e^{-z} \, dz, \quad Re(x) > 0$$

Example 4.12 (Beta Random Variable). *The pdf associated with random variable $X \sim \mathrm{Beta}(\alpha, \beta)$, for positive reals α, β, is defined*

$$f_X(x) \equiv \frac{x^{\alpha-1}(1 - x)^{\beta-1}}{B(\alpha, \beta)}, \quad \text{where } B(\alpha, \beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

and Γ is the Gamma function.

4.4.1 Summary of Discrete and Continuous Random Variables

To summarize, once we have a random variable $X : \Omega \to \mathbb{R}$, we can throw away the sample space and work in $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ with the induced measure $\mathbb{P}_X$.

1. If X is discrete, then let there be some at most countable set $E = \{e_i\}$ where $P(E) = 1$. it turns out that $\mathbb{P}_X$ can be completely defined by a probability mass function $p_X : \mathbb{R} \to \mathbb{R}$ defined

$$p_X(x) = \mathbb{P}_X(\{x\}).$$

Given that we have this pmf , we can define $\mathbb{P}_X$ as such: Given any Borel $B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_X(B) = \sum_{x \in E \cap B} p_X(x)$$

2. If X is continuous, then the Radon-Nikodym Theorem asserts the existence of a nonnegative probability density function f_X that completely describes the probability law $\mathbb{P}_X$. Given that we have this pdf, we can then define $\mathbb{P}_X$ as such: Given any Borel $B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_X(B) = \int_B f_X \, d\lambda$$

The function that characterizes the probability measure $\mathbb{P}_X$, whether it's the pmf p_X or the pdf f_X, is called the **probability distribution** of X.

Theorem 4.4 (σ-Algebra Generated by Random Variable X). *Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X, P_X is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Now, it turns out that the collection of all preimages of Borel sets under X forms a σ-algebra on Ω. We call it*

$$\sigma(X) := \big\{ A \subset \Omega \mid A = X^{-1}(B) \text{ for some } B \in \mathcal{B}(\mathbb{R}) \big\}$$

which is a σ-algebra of Ω. Since X is a measurable function, every $X^{-1}(B)$ is $\mathcal{F}$-measurable, and so $\sigma(X)$ is a sub-σ-algebra of $\mathcal{F}$.

This allows us to "simplify" the σ-algebra $\mathcal{F}$ to the scope of the random variable. That is, let Ω be the sample space of all trajectories of a coin flip before it comes to rest. If we are just looking at whether it is heads or tails, we can define X to have image $\{0, 1\}$. Then, $\sigma(X)$ will be a sub-σ-algebra of $\mathcal{F}$ that looks at only the four subsets $\emptyset, \Omega$, the set of all trajectories landing heads, and the set of all trajectories landing tails. This simplifies $\mathcal{F}$ to a scope that we are interested in.

5 Several Random Variables

Now when we consider several random variables, they will all be defined on the same probability space. Given two random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$, they will each induce a probability law $\mathbb{P}_X$ and $\mathbb{P}_Y$ which completely characterizes them. Note that it is the same underlying randomness that is feeding these random variables, and so if I know some information about the value of X, then we know something about outcome ω, which can be used to find something about the value of Y. To capture this, we can imagine the map $(X, Y) : \Omega \longrightarrow \mathbb{R}^2$ defined $(X, Y)(\omega) := (X(\omega), Y(\omega))$. And just like how X induces a measure P_X onto $\mathbb{R}$, we can imagine (X, Y) inducing a measure onto $\mathcal{B}(\mathbb{R}^2)$, which can be generated by all semi-infinite rectangles $(-\infty, x] \times (-\infty, y]$. Ideally, we would want to put a measure $\mathbb{P}_{X,Y}$ on $\mathbb{R}^2$ s.t.

$$\mathbb{P}_{X,Y}(B) := \mathbb{P}((X, Y)^{-1}(B))$$

where $(X, Y)^{-1}(B) = \{\omega \in \Omega \mid (X(\omega), Y(\omega)) \in B\}$ denotes the preimage of (X, Y). But is $(X, Y)^{-1}(B)$ $\mathcal{F}$-measurable? It turns out that it is, and so now I have a probability law $\mathbb{P}_{X,Y}$ on all Borel sets of $\mathbb{R}^2$.

Definition 5.1 (Joint Probability Law). Given two random variables X, Y on $(\Omega, \mathcal{F}, \mathbb{P})$, the **joint random variable** $(X, Y) : \Omega \longrightarrow \mathbb{R}^2$ is a measurable function defined

$$(X, Y)(\omega) := (X(\omega), Y(\omega))$$

which induces a **joint probability law** $\mathbb{P}_{X,Y} : \mathcal{B}(\mathbb{R}^2) \longrightarrow [0, 1]$ defined

$$\mathbb{P}_{X,Y}(B) := \mathbb{P}((X, Y)^{-1}(B)) \ \forall B \in \mathcal{B}(\mathbb{R})$$

of X, Y. This law captures everything there is about the interdependence of X and Y.

Given joint probability law $\mathbb{P}_{X,Y}$, we can get the probability laws of X and Y separately, but knowing $\mathbb{P}_X$ and $\mathbb{P}_Y$ is not enough to know the joint $\mathbb{P}_{X,Y}$.

Definition 5.2 (Marginal Probability Law)**.** Given a joint probability law $\mathbb{P}_{X,Y}$ of X, Y, we can get the **marginal probability law** of X by feeding in Borel sets of form $B \times \mathbb{R} \in \mathcal{B}(\mathbb{R}^2)$.

$$\mathbb{P}_X(B) = \mathbb{P}_{X,Y}(B \times \mathbb{R})$$

and the marginal probability law of Y as

$$\mathbb{P}_Y(B) = \mathbb{P}_{X,Y}(\mathbb{R} \times B)$$

Definition 5.3 (Joint Cumulative Distribution Function)**.** Since sets of the form $(-\infty, x] \times (-\infty, y]$ are Borel in $\mathbb{R}^2$, the **joint cumulative distribution function**

$$\begin{aligned}
F_{X,Y} &:= \mathbb{P}_{X,Y}\big((-\infty, x] \times (-\infty, y]\big) \\
&= \mathbb{P}\big(\{\omega \mid X(\omega) \le x\} \cap \{\omega \mid Y(\omega) \le y\}\big)
\end{aligned}$$

is well-defined. By abuse of notation, we will write $F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y)$. The marginal CDFs are defined

$$\begin{aligned}
F_X(x) &:= \mathbb{P}_{X,Y}((-\infty, x) \times \mathbb{R}) \\
F_Y(y) &:= \mathbb{P}_{X,Y}(\mathbb{R} \times (-\infty, y))
\end{aligned}$$

Lemma 5.1 (Properties of Joint CDF)**.** *Some common properties of the joint CDF are as follows:*

1. Limits.

$$\lim_{(x,y) \to (+\infty, +\infty)} F_{X,Y}(x,y) = 1 \text{ and } \lim_{(x,y) \to (-\infty, -\infty)} F_{X,Y}(x,y) = 0$$

2. Monotonicity.

$$x_1 \le x_2, \ y_1 \le y_2 \implies F_{X,Y}(x_1, y_1) \le F_{X,Y}(x_2, y_2)$$

3. Continuity from above.

$$\lim_{\epsilon \to 0^+} F_{X,Y}(x + \epsilon, y + \epsilon) = F_{X,Y}(x,y) \text{ for all } x, y \in \mathbb{R}$$

4. Maringal CDFs.

$$\lim_{y \to \infty} F_{X,Y}(x,y) = F_X(x), \quad \lim_{x \to \infty} F_{X,Y}(x,y) = F_Y(y)$$

5.1 Independent Random Variables

Definition 5.4 (Independent Random Variables)**.** Two random variables X, Y are **independent** if $\sigma(X)$ and $\sigma(Y)$ are independent σ-algebras. That is, for any Borel sets $B_1, B_2 \in \mathcal{B}(\mathbb{R})$, the events $X^{-1}(B_1)$ and $Y^{-1}(B_2)$ are independent:

$$\mathbb{P}\big[X^{-1}(B_1) \cap Y^{-1}(B_2)\big] = \mathbb{P}(X^{-1}(B_1)) \, \mathbb{P}(Y^{-1}(B_2))$$

or by abusing notation,

$$\mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}(X \in B_1) \, \mathbb{P}(Y \in B_2)$$

If X, Y are independent, then we can say something about the CDFs

$$F_{X,Y}(x,y) = F_X(x) \, F_Y(y)$$

In fact, we can say something stronger.

Theorem 5.2. *X and Y are independent RVs if and only if*

$$F_{X,Y}(x,y) = F_X(x) \, F_Y(y)$$

Moving onto multiple variables, we can define that $X_1, X_2, \ldots, X_n$ are independent RVs if $\sigma(X_1), \ldots, \sigma(X_n)$ are independent σ-algebras.

5.2   Joint Discrete Random Variables

Definition 5.5 (Joint PMF)**.** Given discrete random variables X and Y, let their countable images be denoted $E_X, E_Y \subset \mathbb{R}$. Then, $E_X \times E_Y$ is also countable, and so the joint random variable (X, Y) is also discrete. This means that we can write for some Borel B of $\mathbb{R}^2$,

$$\mathbb{P}_{X,Y}(B) = \sum_{(x,y) \in (E_X \times E_Y) \cap B} \mathbb{P}_{X,Y}(\{(x,y)\})$$

and we can define the PMF as $p_{X,Y}(x,y) := \mathbb{P}_{X,Y}(\{(x,y)\})$. By abuse of notation, we write $p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$ and write

$$\mathbb{P}_{X,Y}(B) = \sum_{(x,y) \in (E_X \times E_Y) \cap B} \mathbb{P}(X = x, Y = y)$$

If you give me a joint pmf $p_{X,Y}$, by the definition above this determines the entire probability law of $\mathbb{P}_{X,Y}$.

Definition 5.6 (Conditional PMF)**.** Let X, Y be discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The **conditional pmf** of X given $Y = y$ is defined

$$p_{X|Y}(x \mid y) := \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{\mathbb{P}_{X,Y}(\{x,y\})}{\mathbb{P}_Y(\{y\})}$$

and again by abuse of notation, we can simply write

$$\mathbb{P}(X = x \mid Y = y) := \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

Theorem 5.3 (TFAE)**.** *Let X and Y be discrete random variables. Then, the following are equivalent:*

1. *X and Y are independent.*

2. *For all $x, y \in \mathbb{R}$, the events $\{X = x\}$ (aka $X^{-1}(\{x\})$) and $\{Y = y\}$ (aka $Y^{-1}(\{y\})$) are independent. That is,*
$$\mathbb{P}\big[X^{-1}(\{x\}) \cap Y^{-1}(\{y\})\big] = \mathbb{P}(X^{-1}(\{x\})) \, \mathbb{P}(Y^{-1}(\{y\}))$$

3. *For all $x, y \in \mathbb{R}$, $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$.*

4. *For all $x, y \in \mathbb{R}$ such that $p_Y(y) > 0$, we have $p_{X|Y}(x \mid y) = p_X(x)$.*

5.3   Joint Continuous Random Variables

Definition 5.7. X and Y are jointly continuous if the joint law $\mathbb{P}_{X,Y}$ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^2$ (i.e. a Borel set of Lebesgue measure 0 must have $P_{X,Y} = 0$ also).

However, X and Y continuous does not always imply that (X, Y) are jointly continuous! If we have $X \sim \mathcal{N}(0,1)$ and $Y = 2X \sim \mathcal{N}(0,4)$. Jointly continuous allows us to define a pdf on it.

Theorem 5.4 (Radon-Nikodym Theorem)**.** *If X and Y are jointly continuous RVs, then there exists a measurable $f_{X,Y} : \mathbb{R}^2 \longrightarrow [0, \infty)$ s.t. for any $B \in \mathcal{B}(\mathbb{R}^2)$, we have*

$$\mathbb{P}_{X,Y}(B) = \int_B f_{X,Y} \, d\lambda$$

The Radon-Nikodym Theorem guarantees the existence of such $f_{X,Y}$. Taking $B = (-\infty, x] \times (-\infty, y]$, we can define the joint CDF as

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y) := \mathbb{P}_{X,Y}\big((-\infty, x] \times (-\infty, y]\big) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t) \, dt \, ds$$

5.4   Sums of Random Variables

Now given two random variables $X, Y : \Omega \to \mathbb{R}$ that each push their own probability laws $\mathbb{P}_X, \mathbb{P}_Y$ onto $\mathbb{R}$, their sum $Z = X + Y$ is also a random variable that pushes its own probability law $\mathbb{P}_Z$. But the probability law $\mathbb{P}_Z$ can be confusing to define, since given some Borel set $B \in \mathcal{B}(\mathbb{R})$, we must now look at the preimage under the *sum* $X + Y$. A simpler way to approach this is to consider the joint distribution X, Y and look at its distribution.

Definition 5.8 (Distributions of Sums of Random Variables)**.** Take two random variables X, Y and their sum $Z = X + Y$

1. For discrete X, Y, let its joint pmf be defined by $p_{X,Y}(x, y)$. Then, we can simply see that

$$p_Z(z) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, z - x)$$

   which by abuse of notation, we denote

$$\mathbb{P}(Z = z) = \sum_{x \in \mathcal{X}} \mathbb{P}(X = x, Y = z - x)$$

2. For continuous X, Y, let its joint pdf be $f_{X,Y}(x, y)$. Then, we can write

$$f_Z = \int_{\mathbb{R}} f_{X,Y}(x, z - x) \, dx$$

This has a nice visualization, since the joint distribution of X and Y over $\mathbb{R}^2$ is being "summed up/integrated" over the diagonals of $\mathbb{R}^2$, i.e. the lines where $x + y = z$ for some z, sort of like marginalizing over these diagonals. This creates a new "diagonally marginal distribution" Z. Now if X and Y are independent, then their joint distribution is the product of their singular distributions, and so we have

1. for discrete X, Y

$$\mathbb{P}(Z = z) = \sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \, \mathbb{P}(Y = z - x)$$

   The new PMF p_Z is called the convolution of p_X and p_Y, denoted

$$p_Z = p_X * p_Y$$

2. for continuous X, Y

$$f_Z = \int_{\mathbb{R}} f_X(x) \, f_Y(z - x) \, dx$$

   This new PDF f_Z is called the convolution of f_X and f_Y, denoted

$$f_Z = f_X * f_Y$$

Definition 5.9 (Convolution)**.** Given two functions $f, g : \mathbb{R} \longrightarrow \mathbb{R}$, the **convolution** of f and g is a new function $f * g$ defined

$$(f * g)(t) := \int_{\mathbb{R}} f(t) \, g(t - \tau) \, d\tau$$

Theorem 5.5 (Sums of Discrete Variables)**.** *Assume that X and Y are independent.*

1. *$X \sim Binomial(n, p)$, $Y \sim Binomial(m, p) \implies X + Y \sim Binomial(n + m, p)$.*

2. *$X \sim Poisson(\lambda)$, $Y \sim Poisson(\gamma) \implies X + Y \sim Poisson(\lambda + \gamma)$.*

3. *If $X_1, ..., X_n$ are $Geometric(p)$, then $X_1 + ... + X_n$ is $NB(n, p)$.*

Theorem 5.6 (Sums of Densities)**.** *Assume that X and Y are independent.*

1. *$X \sim Normal(\mu_1, \sigma_1^2)$, $Y \sim Normal(\mu_2, \sigma_2^2) \implies X + Y \sim Normal(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

2. *If $X_1, X_2, ..., X_n$ are $Exponential(\lambda)$, then $X_1 + ... + X_n \sim Gamma(n, \lambda)$.*

3. *$X \sim Gamma(n, \lambda)$, $Y \sim Gamma(m, \lambda) \implies X + Y \sim Gamma(n + m, \lambda)$.*

4. *$X \sim Gamma(n, \lambda)$, $Y \sim Exponential(\lambda) \implies X + Y \sim Gamma(n + 1, \lambda)$.*

6   Transformation of Random Variables

7   Expectation and Variance

Recall that given a measure space $(X, \mathcal{A}, \mu)$ and a function $f : X \longrightarrow \mathbb{R}$, we can take some $A \in \mathcal{A}$ and integrate f over it with the Lebesgue integral

$$\int_A f \, d\mu$$

If $A = X$, then it is conventional not to write it in the integral at all (so it looks like an indefinite integral).

Definition 7.1 (Expectation). Given a probability space $(\Omega, \mathcal{F}, \mathbb{P}$ and a random variable $X : \Omega \longrightarrow \mathbb{R}$, the **expectation** of X is defined

$$\mathbb{E}[X] := \int_\Omega X \, d\mathbb{P}$$

Definition 7.2 (Expectation of Discrete RV). If X is a discrete random variable *that takes positive values*, then let $E = \{e_1, e_2, \ldots\}$ denote the set where $\mathbb{P}_X(E) = 1$, and let $E_i = X^{-1}(\{e_i\}) \subset \Omega$. Then, we can see that since X is constantly e_i on E_i,

$$\int_{E_i} X \, d\mathbb{P} = e_i \cdot \mathbb{P}(E_i) = e_i \cdot \mathbb{P}_X(\{e_i\}) = e_i \cdot \mathbb{P}(X = e_i)$$

which implies

$$\mathbb{E}[X] = \int_\Omega X \, d\mathbb{P} = \sum_{i=1}^\infty \int_{E_i} X \, d\mathbb{P} = \sum_{i=1}^\infty e_i \cdot \mathbb{P}(X = e_i)$$

If X is discrete RV possibly taking negative values, then let $X = X^+ - X^-$, where $X^+ = \max(X, 0)$ and $X^- = -\min(X, 0)$. Then, we can compute

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

which is well-defined as long as we don't have "$\infty - \infty$."

Note that the reason why expectations of the form $\infty - \infty$ are indeterminate is because of the Riemann rearrangement theorem.

Theorem 7.1 (Riemann's Rearragenement Theorem). *Given a series $\sum a_n$ that is conditionally convergent (i.e. converges but not absolutely convergent), the terms can be arranged so that the new series converges to an arbitrary real number, or diverges.*

Lemma 7.2 (Properties of Expectation). *Let X and Y be random variables with finite expectations.*

1. *Monotonicity: If $g(x) \geq h(x)$ for all $x \in \mathbb{R}$, then*

$$\mathbb{E}[g(X)] \geq \mathbb{E}[h(X)]$$

2. *Linearity: For all $a, b, c \in \mathbb{R}$,*

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

We now show a widely-used, but nontrivial, theorem.

Theorem 7.3 (Expectation of Independent Events). *Given independent RVs X and Y,*

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$

Proof. We show only for simple random variables which will give us a start in proving for all random variables in full generality. Let X and Y be simple random variables, i.e.

$$X = \sum_i a_i \mathbb{I}_{A_i} \text{ and } Y = \sum_j b_j \mathbb{I}_{B_j}$$

Since $\{A_i\}_i$ and $\{B_j\}_j$ are both partitions, $\{A_i \cap B_j\}_{i,j}$ is also a partition, and

$$XY = \sum_{i,j} a_i b_j \, \mathbb{I}_{A_i \cap B_j}$$

Its expectation can be expanded out by linearity, and since $\mathbb{E}[\mathbb{I}_A] = \mathbb{P}(A)$, we have

$$\mathbb{E}[XY] = \sum_{i,j} a_i b_j \, \mathbb{P}(A_i \cap B_j)$$

$$= \sum_{i,j} a_i b_j \, \mathbb{P}(A_i) \, \mathbb{P}(B_j) = \mathbb{E}[X] \, \mathbb{E}[Y]$$

Now that we have proved for simple random variables, we can just approximate X from below using simple functions. ∎

Theorem 7.4 (Tail Sum Formula). *If a discrete random variable X takes values in the non-negative integers $\{0, 1, 2, 3, ...\}$, then*

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

In any case (continuous or discrete), if X is a non-negative random variable, then

$$\mathbb{E}(X) = \int_0^{\infty} \mathbb{P}(X > x) \, dx = \int_0^{\infty} 1 - F(x) \, dx$$

where F is the CDF of X.

Proof. Suppose that X takes values in $\{0, 1, 2, 3, ...\}$. Then,

$$\mathbb{E}(X) = \sum_{k \geq 1} k \, \mathbb{P}(X = k)$$

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

$$= \sum_{k \geq 1} \sum_{j=1}^{k} \mathbb{I}_{j \leq k} \, \mathbb{P}(X = k)$$

$$= \sum_{j=1}^{\infty} \sum_{k \geq 1} \mathbb{I}_{j \leq k} \, \mathbb{P}(X = k)$$

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

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

∎

Corollary 7.4.1. *For any $m > 0$ and $\alpha > 0$,*

$$\mathbb{P}(|X| > \alpha) \leq \frac{1}{\alpha^m} \mathbb{E}(|X|^m)$$

Example 7.1 (Geometric RV). *Recall that given $X \sim \mathrm{Geometric}(p)$, we have $\mathbb{P}(X = i) = (1-p)^{i-1} p$ for $i \geq 1$. So,*

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x \, \mathbb{P}(X = x) = \sum_{x=1}^{\infty} x \, (1-p)^{i-1} p = \frac{p}{(1 - (1-p))^2} = \frac{1}{p}$$

Example 7.2 (Infinite Expectation). *Let us have discrete random variable s.t. $\mathbb{P}(X = k) = \frac{6}{\pi^2} \frac{1}{k^2}$ for $k \geq 1$. So,*

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \, \mathbb{P}(X = k) = \frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k} = +\infty$$

Example 7.3 (Undefined Expectation). *Let $\mathbb{P}(X = k) = \frac{3}{\pi^2}\frac{1}{k^2}$ for $k \in \mathbb{Z} \setminus \{0\}$. The expectation of this can be computed by getting the expectation of all the positive terms and the negative terms.*

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] = \sum_{k=1}^{\infty} k \cdot \frac{3}{\pi^2}\frac{1}{k^2} + \sum_{k=1}^{\infty} (-k) \cdot \frac{3}{\pi^2}\frac{1}{k^2} = \infty - \infty$$

Note that by the Riemann rearrangement theorem, we can't just say that the expectation is 0 since the terms "cancel out." We could only do this if the series is absolutely convergent also, which works if X takes positive values only.

Note that when we compute expectation, what we do it multiply the pmf/pdf by x and sum/integrate over it. The Cauchy distribution is a power function of form $\frac{1}{x^2}$, so if we multiply it by x, we have the new $\frac{1}{x}$ which is harmonic and therefore divergent.

7.1 Law of the Unconscious Statistician

Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, this induces a probability law $\mathbb{P}_X$ acting as a measure on $\mathbb{R}$. Assume that this probability law $\mathbb{P}_X$ is known. Now introduce a function $g : \mathbb{R} \to \mathbb{R}$. We can create a new random variable $Y = g \circ X : \Omega \to \mathbb{R}$ with its own probability law $\mathbb{P}_Y$ on $\mathbb{R}$. Since we already know the probability distribution of X, so we can easily get the expected value of X as (in the discrete case)

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

where $\mathcal{X}$ is the support of X. But what if we wanted to get the expected value of Y?

$$\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} y \cdot \mathbb{P}(Y = y) = ?$$

The problem is that we don't know the probability distribution of Y. But since we know that all the values of X are transformed by g, we are taught to compute it in terms of the probability distribution of X.

$$\mathbb{E}[Y] = \sum_{x \in \mathcal{X}} g(x) \cdot \mathbb{P}(X = x)$$

This "identity" that is often used must actually be treated as a rigorous theorem.

Theorem 7.5 (LOTUS). *Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable $X : \Omega \to \mathbb{R}$, and a function $g : \mathbb{R} \to \mathbb{R}$, the expectation of $g(X)$ is*

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g \, d\mathbb{P}_X$$

1. For the discrete case, the above integral simplifies to

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x)$$

2. For the continuous case, we have

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g \, f_X \, d\lambda$$

and in particular,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, f_X \, d\lambda$$

Example 7.4 (Expectation of Exponential RV). *The pdf of exponential random variable X is defined $f_X = ke^{-kx}$ for $x \geq 0$. So,*

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X \, d\lambda = \int_0^{\infty} xke^{-kx} \, dx = \frac{1}{k}$$

Similarly, if we want the expectation of X^2, then we can get

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f_X \, d\lambda = \int_0^{\infty} x^2 ke^{-kx} \, dx = \frac{2}{k^2}$$

Example 7.5 (Expectation of Gaussian RV). *The expectation of a Gaussian random variable X is*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \mu$$

Example 7.6 (Expectation of One-Sided Cauchy). *If we have $f_X(x) = \frac{2}{\pi}\frac{1}{1+x^2}$ for $x \geq 0$, then*

$$\mathbb{E}[X] = \int_0^{\infty} \frac{2}{\pi}\frac{x}{1+x^2} \, dx$$

and making the substitution $t = \frac{1+x^2}{,} \, dt = 2x$, we have

$$\int_1^{\infty} \frac{1}{\pi}\frac{1}{t} \, dt = \frac{\ln(t)}{\pi}\bigg|_1^{\infty} = +\infty$$

Example 7.7 (Expectation of Two-Sided Cauchy). *The two-sided Cauchy is just another copy of the one sided into the negatives, so $f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ for $x \in \mathbb{R}$. The expectation of X should be split up into for positive and negative images, but computing it gives*

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] = \int_0^{\infty} \frac{1}{\pi}\frac{x}{1+x^2} \, dx - \int_{-\infty}^0 \frac{1}{\pi}\frac{x}{1+x^2} \, dx = \infty - \infty$$

and so it is undefined.

7.2   Variance

Definition 7.3 (Variance). Let X be a random variable and suppose $\mathbb{E}[X] < \infty$. The **variance** of X is defined

$$\mathrm{Var}[X] = \sigma_X^2 := \mathbb{E}[(X - \mathbb{E}[X])^2]$$

and $\sigma_X = \sqrt{\mathrm{Var}[X]}$ is called the **standard deviation**. This is a measure of how much the probability distribution deviates from its mean. We can use linearity of expectation to write

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}\big[X^2 + \mathbb{E}[X]^2 - 2X\mathbb{E}[X]\big] \\
&= \mathbb{E}[X^2] + \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[X] \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2
\end{aligned}$$

which is often easier to compute, since it only requires us to compute the expectation of X and X^2. Since variance is always nonnegative, we also know that $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$. The variance is always defined, whether it's finite or $+\infty$.

Proposition 7.6. *The covariance of a random variable X is 0 if and only if it constant almost everywhere on Ω.*

Proof. The if part is easy, so let's prove the only if part. Let $\mathbb{E}[(X - \mathbb{E}(X))^2] = 0$. Then, we can think of the function $x \mapsto (x - \mathbb{E}(X))^2$ and write the variance as

$$\mathrm{Var}[X] = \int_{\Omega} (X - \mathbb{E}(X))^2 \, d\mathbb{P} = 0$$

But by nonnegativity of the function, we know that $(X - \mathbb{E}[X])^2 = 0$ w/ probability 1, which implies that $X = \mathbb{E}[X]$ with prob. 1. ∎

Lemma 7.7 (Properties of Variance). *Let X and Y be random variables with well-defined variances.*

1. *Translation Invariance: Given that $X + a$ is a new random variable defined $(X+a)(\omega) = X(\omega)+a$,*

$$\mathrm{Var}[X] = \mathrm{Var}[X + a]$$

2. *Quadratic Scaling: Given that aX is a new random variable defined $(aX)(\omega) = a\,X(\omega)$,*

$$\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$$

From the properties of expectation and variance, we can now **standardize** a random variable X. If X is a random variable with mean $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \text{Var}(X)$, then the random variable

$$Y = \frac{X - \mu}{\sigma}$$

has mean $\mathbb{E}(Y) = 0$ and variance $\text{Var}(Y) = 1$.

Example 7.8 (Bernoulli). *Given $X \sim \text{Bernoulli}(p)$, we have*

$$\mathbb{E}[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = p$$
$$\mathbb{E}[X^2] = 0^2 \cdot \mathbb{P}(X = 0) + 1^2 \cdot \mathbb{P}(X = 1) = p$$

and so $\text{Var}[X] = p - p^2 = p(1 - p)$.

Example 7.9 (Poisson). *Given $X \sim \text{Poisson}(X)$, then*

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

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k^2 \cdot \frac{e^{-\lambda} \lambda^k}{k!} = \ldots = \lambda^2 + \lambda$$

So $\text{Var}[X] = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Example 7.10 (Uniform). *Let $X \sim \text{Uniform}[a, b]$. Then,*

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X \, d\lambda = \int_a^b x \cdot \frac{1}{b-a} \, dx = \frac{a+b}{2}$$

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f_X \, d\lambda = \int_a^b \frac{x^2}{b-a} \, dx = \frac{a^2 + ab + b^2}{3}$$

So $\text{Var}[X] = \ldots = \frac{1}{12}(b-a)^2$. *This is consistent with the fact that if we spread out our measure over a wider interval, then the variance will be bigger.*

Example 7.11 (Exponential). *Let $X \sim \text{Exp}(\lambda)$. Then, $\mathbb{E}[X] = \frac{1}{\lambda}$ and $\mathbb{E}[X^2] = \frac{2}{\lambda^2}$, so*

$$\text{Var}[X] = \frac{1}{\lambda^2}$$

This is consistent with the fact that if λ is greater, then the pdf is much more concentrated at 0, making the variance small.

Just like how we explained that computing finiteness or infiniteness of expectation is similar to multiplying the pmf/pdf by x and determining if the series/integral converges or diverges, we can do the same for variance by multiplying the pmf/pdf by x^2. For a probability distribution of form $\frac{1}{x^2}$, it diverges if we multiply by x and also diverges if we multiply by x^2. But also, we could construct a distribution where the expectation may be finite, but the variance may be infinite. For example, if we have a distribution of form $\frac{1}{x^3}$, multiplying it by x leads to form $\frac{1}{x^2}$, which is finite (so finite expectation), but multiplying by x^2 leads to a harmonic, i.e. infinite variance.

7.3   Covariance

The variance is a measure for one random variable X, which measures how much it deviates from its mean. Now, the covariance is defined for two random variables and captures how they jointly vary.

Definition 7.4 (Covariance). The **covariance** of random variables X and Y is defined as

$$\text{Cov}[X, Y] = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[X])\big]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

where the intermediate expectations are well-defined. X and Y are said to be **uncorrelated** if

$$\text{Cov}[X, Y] = 0$$

The covariance is also easy to interpret. Given two random variables X and Y, if whenever X is greater than its expected value $\mathbb{E}[X]$, Y also tends to be greater than $\mathbb{E}[Y]$, then the covariance will be some positive number. If they tend to be on opposite sides of their expected values, then the covariance will be negative. And the degree with which these RVs lie on which side of the expected value determines the magnitude of the covariance.

Theorem 7.8. *If X and Y are independent random variables, then they are uncorrelated, meaning that independence is a stronger condition.*

We show an example of why the converse is not true. Consider $X \sim \mathrm{Uniform}[-1, 1]$. We can show that x and $Y = X^2$ are dependent but uncorrelated. It is clearly dependent, but its covariance is

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$
$$= \mathbb{E}[X^3] - \mathbb{E}[X]\,\mathbb{E}[X^2]$$
$$= \int_{-1}^{1} x^3 \cdot 1 \, dx - 0 \cdot \mathbb{E}[X^2] = 0$$

Theorem 7.9. *If X and Y are two random variables, then*

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

and by induction, we can show that

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i) + \sum_{i,j} \mathrm{Cov}(X_i, X_j)$$

Proof. Simple computation. The LHS expands to

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

■

7.4   Correlation, Hilbert Space of Random Variables

Definition 7.5 (Correlation Coefficient). The **correlation coefficient** of random variables X and Y is defined

$$\rho_{X,Y} = \mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

In some sense the correlation is a scaled version of the covariance. It is scale-invariant, and it is always a number that lies between -1 and 1, making it a nice way to represent the correlation between two variables without having to worry about scale. We can prove this.

Theorem 7.10 (Cauchy-Schwartz). *For any two random variables X, Y, we have $|\mathrm{Cov}(X, Y)| \leq \sigma_X \sigma_Y$, or in other words,*

$$-1 \leq \rho_{X,Y} \leq 1$$

Furthermore, whenever $\rho_{X,Y} = 1$ or -1, there exists a deterministic relationship between X and Y.

1. If $\rho_{X,Y} = 1$, there exists a $a > 0$ s.t.

$$Y - \mathbb{E}[Y] = a(X - \mathbb{E}[X])$$

2. If $\rho_{X,Y} = -1$ there exists a $a < 0$ s.t.

$$Y - \mathbb{E}[Y] = a(X - \mathbb{E}[X])$$

This implies that $\mathrm{Corr}(X, Y) = \pm 1$ indicates that the joint distribution of (X, Y) is concentrated on a line in $\mathbb{R}^2$.

The fact that this is called the Cauchy-Schwartz inequality hints at the existence of inner products, norms, and vector spaces. That is, we can treat the random variables X, Y as vectors in the functional space of real-valued maps over Ω. In some sense, $\text{Cov}(X, Y)$ sort-of plays the role of an inner product.

1. It satisfies symmetricity:
$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[YX] - \mathbb{E}[Y]\mathbb{E}[X] = \text{Cov}(Y, X)$$

2. It satisfies binlinearity. It suffices to show only for first argument, since we have symmetricity.
$$\begin{aligned}
\text{Cov}(aX + bY, Z) &= \mathbb{E}[(aX + bY)Z] - \mathbb{E}[aX + bY]\mathbb{E}[Z] \\
&= a\mathbb{E}[XZ] + b\mathbb{E}[YZ] - a\mathbb{E}[X]\mathbb{E}[Z] - b\mathbb{E}[Y]\mathbb{E}[Z] \\
&= a\big(\mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z]\big) + a\big(\mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z]\big) \\
&= a\,\text{Cov}(X, Z) + b\,\text{Cov}(Y, Z)
\end{aligned}$$

3. We want the inner product of X with itself to always be greater than 0, with equality holding iff $X = 0$. Indeed, we have
$$\text{Cov}(X, X) = \text{Var}(X) \geq 0$$
but it is not necessarily true that $\text{Var}(X) = 0 \implies X = 0$. We can say that X is equal to a constant almost everywhere at best. We can solve this problem by looking at the functional subspace of 0-mean random variables (which is a vector space due to linearity of expectation). So now all random variables X that are 0 almost everywhere have inner product 0, so we must add an equivalence class on this subspace that says two X, Y are equivalent if they agree almost everywhere.

The standard deviation σ_X and σ_Y act as norms on this quotient subspace of 0-mean random variables. So the correlation coefficient $\rho_{X,Y}$ can be interpreted as the cosine of the angle between X and Y. This now makes our desired space a Hilbert space, and our uncorrelated random variables are like orthogonal vectors.

7.5 Conditional Expectation

Given some joint probability distribution, we can define the conditional expectation as the expectation of a conditional distribution. This construct is very elementary, and we will start with the discrete case.

Definition 7.6 (Discrete Conditional Expectation)**.** Let X, Y be discrete random variables, with pmf $p_{X,Y}(x, y)$. Recall that the conditional pmf is
$$p_{X|Y}(x \mid y) := \frac{p_{X,Y}(x, y)}{p_Y(y)}$$
The **conditional expectation** of X given $Y = y$ is
$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x\, p_{X|Y}(x \mid y)$$

Now we can treat this conditional expectation as a function, letting $\psi(y) := \mathbb{E}[X \mid Y = y]$. This function ψ assigns a value to every $y \in \mathcal{Y}$, and therefore can be considered a random variable in itself. The random variable $\psi(Y) := \mathbb{E}[X \mid Y]$ is called the **conditional expectation** of X given Y.

Note that $\mathbb{E}[X \mid Y]$ is not a value! It is a random variable. Now for the jointly continuous case.

Definition 7.7 (Joint Conditional Expectation)**.** Let X, Y be jointly continuous with joint pdf $f_{X,Y}(x, y)$. Recall that the conditional pdf is
$$f_{X|Y}(x \mid y) := \frac{f_{X,Y}(x, y)}{f_Y(y)}$$
The **conditional expectation** of X given $Y = y$ is
$$\mathbb{E}[X \mid Y = y] = \int_{x \in \mathbb{R}} x\, f_{X|Y}(x \mid y)\, dx$$

Again, we can set $\psi(y) := \mathbb{E}[X \mid Y = y]$, which is a function of y and therefore a random variable. Therefore the **conditional expectation** of X given Y is $\mathbb{E}[X \mid Y]$, which is a $\sigma(Y)$-measurable random variable.

7.6   Moment Generating Function

Definition 7.8 (Moment). The **nth (raw) moment** of a random variable X is $\mathbb{E}[X^n]$. Unlike the raw moment, which is calculated around the origin, the **nth central moment** of X is its moment centered around its mean $\mathbb{E}[(X - \mathbb{E}[X])^n]$.

1. the first moment is the mean $\mathbb{E}[X]$

2. the second central moment is the variance $\mathbb{E}[(X - \mathbb{E}[X])^2]$

3. the third central moment, divided by σ^3, is the skew $\frac{1}{\sigma^3}\mathbb{E}[(X - \mathbb{E}[X])^3]$

Definition 7.9 (Moment Generating Function (MGF)). The **moment generating function** associated with a random variable X is a function $M_X : \mathbb{R} \longrightarrow [0, \infty]$ defined

$$M_X(s) := \mathbb{E}[e^{sX}]$$

It is like an exponential moment. The region of convergence of M_X is the set

$$D_X = \{s \mid M_X(s) < \infty\}$$

We specify them for discrete and continuous random variables:

1. If X is discrete, with PMF $p_X(x)$, then by LOTUS we have

$$M_X(s) = \sum_{x \in \mathcal{X}} e^{sx} p_X(x)$$

2. If X is continuous, with PDF $f_X(x)$, then

$$M_X(s) = \int_{\mathbb{R}} e^{sx} f_X(x) \, dx$$

Note that the MGF is really a complex-valued function, and results about it require a fair amount of complex analysis, but we will restrict ourselves to the real-values.

Example 7.12 (Exponential RV). *The pdf of $X \sim \text{Exponential}(\mu)$ is $f_X(x) = \mu e^{-\mu x}$ for $x \geq 0$. The MGF is*
$$M_X(s) := \int_0^\infty \mu e^{-\mu x} e^{sx} \, dx = \frac{\mu}{\mu - s} \text{ for } s < \mu$$
and if $s > \mu$, then $e^{(s-\mu)x}$ increases and the MGF is ∞.

Example 7.13 (Gaussian RV). *The pdf of a standard Gaussian X is $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for $x \in \mathbb{R}$, and the MGF is*

$$M_X(s) = \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{sx} \, dx = e^{s^2/2} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-s)^2}{2}} \, dx = e^{s^2/2}$$

which is valid for all $s \in \mathbb{R}$.

Example 7.14 (Cauchy RV). *If we have $f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}$ for $x \in \mathbb{R}$, the MGF is*

$$M_X(s) = \int_{-\infty}^\infty \frac{e^{sx}}{\pi(1 + x^2)} \, dx = \begin{cases} 1 & \text{if } s = 0 \\ \infty & \text{if } s > 0 \\ \infty & \text{if } s < 0 \end{cases}$$

So the region of convergence is just $\{0\}$. It is infinity everywhere else since the exponential function grows exponentially as $x \to \pm\infty$.

Theorem 7.11. *Suppose $M_X(s)$ is finite for all $s \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$. Then, M_X uniquely determines the CDF of X.*

8   Concentration Inequalities

Concentration inequalities give you probability bounds on random variables taking atypical values. For example, given a random variable with certain mean and variance, the probability of that random variable taking values outside a certain range around the mean is very small. It's called concentration because the probability concentrates around a certain range.

The most important concentration inequality is Markov's inequality.

Theorem 8.1 (Markov's Inequality). *If X is a non-negative random variable of finite expectation and $\alpha > 0$, then*

$$\mathbb{P}(X > \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$$

That is, the probability that X takes a value greater than α is at most the expectation of X divided by α. This is meaningful only when $\mathbb{E}[X] < \alpha$, since otherwise the RHS will be greater than 1.

Proof. Given any $\alpha > 0$, we can set

$$X = X \cdot \mathbb{I}_{X \leq \alpha} + X \cdot \mathbb{I}_{X > \alpha}$$

and by linearity,

$$
\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X \cdot \mathbb{I}_{X \leq \alpha} + X \cdot \mathbb{I}_{X > \alpha}] \\
&\geq \mathbb{E}[X \cdot \mathbb{I}_{X > \alpha}] \\
&\geq \alpha \mathbb{E}[\mathbb{I}_{X > \alpha}] \\
&= \alpha \, \mathbb{P}(X > x)
\end{aligned}
$$

∎

In other words, the probability that $X > \alpha$ goes down at least as fast as $1/\alpha$. For example, setting $\alpha = 2\mathbb{E}[X]$, the probability that X takes value that is at least twice its expectation is at most $1/2$. Furthermore, as X gets very large, the probability that it will take a value beyond a large α goes down faster than $1/\alpha$. But this is a very conservative inequality, and usually the probability goes down much faster.

Corollary 8.1.1 (Markov's Inequality, 2nd Form).

$$\mathbb{P}\big(X \geq s \, \mathbb{E}(X)\big) \leq \frac{1}{s}$$

Proof. Let $x = s\mathbb{E}(X)$ in Markov's inequality to get this second form of Markov's inequality.   ∎

Markov's inequality is very conservative but very general, too. If we make further assumptions about the random variable X, we can often make stronger bounds. Chebyshev's inequality assumes a (possibly negative) random variable with finite variance and states that the probability will go down as $1/x^2$.

Theorem 8.2 (Chebyshev Inequality). *Given random variable X, if $\mathbb{E}[X] = \mu < +\infty$ and $\mathrm{Var}(X) = \sigma^2 < +\infty$, then for all $\alpha > 0$,*

$$\mathbb{P}\big(|X - \mu| > k\sigma\big) \leq \frac{1}{k^2}$$

That is, the probability that X takes a value further than k standard deviations away from μ goes down by $1/k^2$. Another way of saying it is

$$\mathbb{P}(|X - \mu| > c) \leq \frac{\sigma_X^2}{c^2}$$

since it talks about just some absolute distance c away from the mean μ, rather than the number of standard deviations k away from μ. But they are essentially the same thing if we let $c k\sigma$. Therefore, if σ is small, then this bound will be small since there is more concentration in the mean.

Proof. We apply Markov's inequality to the non-negative random variable $|X - \mu|^2$.

$$\mathbb{P}(|X - \mu|^2 > k^2\sigma^2) \leq \frac{\mathbb{E}(|X - \mu|^2)}{k^2\sigma^2} = \frac{1}{k^2}$$

since the numerator on the RHS is the definition of variance.   ∎

But even Chebyshev's inequality turns out to be quite loose, and even this $1/k^2$ is not a very nice bound.

Example 8.1. *For the normal distribution, recall the 67-95-99.7 rule. It is well known that the probability of a random variable taking values within 2 standard deviations from the mean is 95%, so the probability that it takes outside is 5%, or 1/20, which is less than the $1/2^2 = 1/4$ bound given by Chebyshev.*

A very important theorem that gives you an exponential bound on $\mathbb{P}(X > \alpha)$.

Theorem 8.3 (Chernoff Bound). *Given a random variable X, assume that its moment generating function $M_X(s) = \mathbb{E}[e^{sX}]$ is finite for every $s \in [-\epsilon, \epsilon]$. Then, since $x \mapsto e^{sx}$ is monotonically increasing, we have the identity*

$$\mathbb{P}(X > \alpha) = \mathbb{P}(e^{sX} > e^{s\alpha}) \text{ for } s > 0$$

But since the new random variable e^{sX} is nonnegative, we can now go back to Markov inequality and write

$$\mathbb{P}(X > \alpha) = \mathbb{P}(e^{sX} > e^{s\alpha}) \geq \frac{\mathbb{E}[e^{sX}]}{e^{s\alpha}} = M_X(s)\, e^{-s\alpha}$$

for $s > 0$ (for identity above to hold) and $s \in D_X$ (and it is in domain of convergence). Now, we have an exponentially decaying bound in terms of α. We have the freedom to choose s, since our bound is in terms of α, so we must choose s that minimizes $M_X(s)\, e^{-s\alpha}$. Ultimately, our best bound is

$$\mathbb{P}(X > \alpha) \leq \inf_{s > 0} M_X(s)\, e^{-s\alpha}$$

After we optimize over s what remains on the RHS is a function of α.

Sometimes, we choose to work with the log moment generating function, so we define

$$\Lambda_X(s) := \log M_X(s)$$

and our bound is

$$\mathbb{P}(X > \alpha) \leq \inf_{s > 0} M_X(s)\, e^{-s\alpha} = \exp\left(-\sup_{s > 0}[sx - \Lambda_X(s)]\right) = e^{-\Lambda^*(\alpha)}$$

where Λ^* is the optimized function in terms of s.

9   Convergence of Random Variables

Definition 9.1 (Sure Convergence of RVs). The sequence of random variables $\{X_n\}_{n \in \mathbb{N}}$ is said to **converge pointwise** or **converge surely** to X if

$$X_n(\omega) \to X(\omega)$$

for every $\omega \in \Omega$.

But this definition is too strong of a form of convergence, since in probability we don't care about values over sets of measure 0. So we never use this.

Definition 9.2 (Almost Sure Convergence of RVs). The sequence of random variables $\{X_n\}_{n \in \mathbb{N}}$ is said to **converge almost surely** or **converge with probability 1** to X if

$$X_n(\omega) \to X(\omega)$$

on a subset of probability 1. That is,

$$\mathbb{P}\big(\{\omega \in \Omega \mid \lim_{n \to \infty} X_n(\omega) = X(\omega)\}\big) = 1$$

It can be shown that this set of ω's can be considered an event in $\mathcal{F}$.

Definition 9.3 (Convergence in Probability). The sequence of random variables $\{X_i\}_{i \in \mathbb{N}}$ is said to **converge to X in probability** if for all $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\big(|X_n - X| > \epsilon\big) = 0$$

Note that while it may not be apparent now, convergence with probability 1 and convergence with probability are very different. Almost sure convergence has the limit inside the probability, which indicates that we are talking about convergence of a sequence of random variables. On the other hand, convergence in probability has the limit on the outside, which talks about convergence of a sequence of probabilities. But a key point is that almost sure convergence implies convergence in probability. It happens so because there could exist a subset of small probability in Ω where the X_n's and X need not be close, but the probabilities of them deviating over whole Ω is small.

9.1   Laws of Large Numbers

9.2   Central Limit Theorem

10   Multivariate Gaussians

Recall $X \sim \mathcal{N}(\mu, \sigma^2)$ implies that its pdf is

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

Now we will consider a Gaussian random *vector*, which can be considered a vector of random variables

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

mapping from Ω to $\mathbb{R}^n$. It is not merely a vector where every X_i is Gaussian, as we will show later. That is, a joint distribution that has all n marginal distributions Gaussians does not make a multivariate Gaussian.

This measurable function $\mathbf{X} : \Omega \to \mathbb{R}^n$ induces a probability law $\mathbb{P}_X$ on $\mathcal{B}(\mathbb{R}^n)$, and the Radon-Nikodym theorem states the existence of a pdf f_X such that $\mathbb{P}_X(B) = \int_B f_X \, d\lambda$.

10.1   Bivariate Gaussians

Let us first begin with two-variable Gaussians.

Definition 10.1 (Standard Bivariate Gaussian RV)**.** A random variable (X, Y) is said to be a **standard bivariate Gaussian** if its pdf is of form

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right) \text{ for } \rho \in (-1, 1)$$

Proposition 10.1. *Given a standard bivariate Gaussian (X, Y),*

1. *X and Y are marginally distributed as $\mathcal{N}(0, 1)$. That is, if we integrate a variable (say, x) out, we will get a univariate standard Gaussian pdf of the other (y):*

$$\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right) dx = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

2. *$\rho_{X,Y}$, the correlation coefficient of X and Y, is equal to ρ.*

3. *The conditional distribution of X given $Y = y$ is $X \mid Y = y \sim \mathcal{N}(\rho y, 1 - \rho^2)$. That is,*

$$f_{X|Y}(x \mid y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(x-(\rho y))^2}{2(1-\rho^2)} \right)$$

4. *From (3), we can see that the conditional expectation $\mathbb{E}[X \mid Y = y] = \rho y$ since $X \mid Y = y$ has mean at ρy. Therefore, the conditional expectation of X given Y (which is a random variable) is*

$$\mathbb{E}[X \mid Y] = \rho Y$$

i.e. $\mathbb{E}[X \mid Y]$ is a linear function of Y.

The formula of the general bivariate Gaussian $\mathbf{X} = (X_1, X_2)$ pdf is messy, but we will put it here.

$$f_{X_1,X_2}(x,y) = \frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{\exp}\left[-\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2}\right)\right]$$

It is cleaner to put it into matrix form.

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}}\exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right)$$

where

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \mathbb{E}\big[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\big] = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{pmatrix}$$

Note that visually, $\boldsymbol{\Sigma}$ will determine how much the Gaussian distribution is "stretched" on one way or another. Obviously, the "peak" of the distribution will be $\boldsymbol{\mu}$. If $\boldsymbol{\Sigma} = I$, then we could visualize the Gaussian distribution as being perfectly symmetric. However, if we scale the distribution up to a certain constant (below shown $\boldsymbol{\Sigma} = I$, $\boldsymbol{\Sigma} = 0.61I$, $\boldsymbol{\Sigma} = 2I$), we get
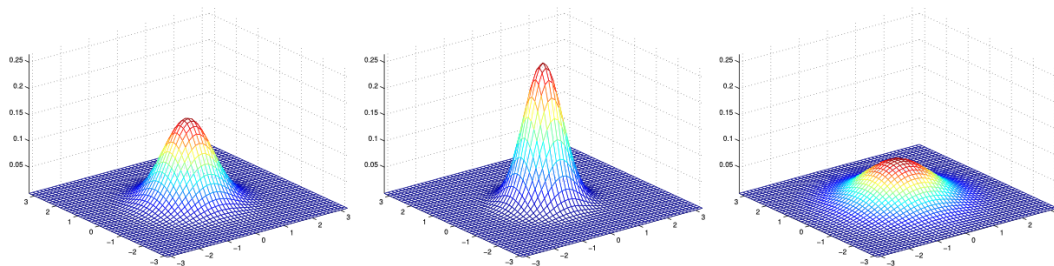
Now we've made a remark before that given a multivariate distribution $\mathbf{X} = (X_1, \ldots, X_n)$, all of its marginal distributions being Gaussian does not mean that $\mathbf{X}$ is a multivariate Gaussian. We give a counterexample.

Example 10.1. *Let Y_1, Y_2 be iid random variables distributed according to the pdf*

$$f_Y(y) = \sqrt{\frac{2}{\pi}}e^{-y^2/2} \text{ for } y > 0$$

which we can interpret as a one-sided Gaussian. Let $W \sim \text{Bernoulli}(\frac{1}{2})$ be independent of Y_1 and Y_2. Now, define the random variables

$$X_1 = W Y_1 \text{ and } X_2 = W Y_2$$

Now note that Y_1 and Y_2 are both positive, and since X_1 and X_2 are both dependent on the same value of W, it is either X_1 and X_2 are both positive or both negative. So, the joint distribution of X_1, X_2 will be on only the 1st and 3rd quadrant with no mass on the 2nd and 4th.
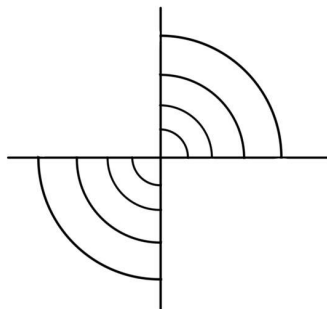
This is clearly not a multivariate Gaussian, even though the marginals are $X_1, X_2 \sim \mathcal{N}(0,1)$. We could make the degenerate case that $X_1 = X_2$, which would make the image of (X_1, X_2) just the line at $x_1 = x_2$, but we can think of this as a degenerate Gaussian with a singular $\boldsymbol{\Sigma}$.

10.2   Multivariate Gaussians

There are three equivalent definitions of multivariate Gaussians of n-variables.

Definition 10.2 (Multivariate Gaussian). Let us have a vector-valued random variable $\mathbf{X} = (X_1 \dots X_n)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. $\mathbf{X}$ is a **multivariate Gaussian distribution** with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and symmetric, positive-definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ if its probability density function is

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(x - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu}) \right)$$

The covariance matrix $\boldsymbol{\Sigma}$ is the $n \times n$ matrix whose (i, j)th entry is $\mathrm{Cov}(X_i, X_j)$. That is, for any random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$, its covariance matrix

$$\boldsymbol{\Sigma} = \mathbb{E}\big[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\big] = \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

is positive definite and symmetric, which implies by the spectral theorem we can break it down into n orthogonal eigenspaces of positive eigenvalues.

2. X is a multivariate Gaussian distribution if it can be expressed as

$$\mathbf{X} = \mathbf{D}\mathbf{w} + \boldsymbol{\mu}$$

where $\mathbf{w}$ is a vector of independent $\mathcal{N}(0, 1)$ Gaussians, $\boldsymbol{\mu} \in \mathbb{R}^n$, and $\mathbf{D} \in \mathbb{R}^{n \times n}$. The mean of $\mathbf{X}$ is $\boldsymbol{\mu}$ and its covariance is $\boldsymbol{\Sigma} = \mathbf{D}\mathbf{D}^T$; $\mathbf{D}$ is called the **standard deviation matrix**. When modeling high-dimensional Gaussians, this way is most computationally feasible.

3. X is a multivariate Gaussian distribution if for every $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{a}^T \mathbf{x}$ is a Gaussian RV. This means that if we take $\mathbf{a} = \mathbf{0}$, then the entire $\mathbf{X}$ is constantly 0, which we will take to be the degenerate Gaussian with mean, variance 0.

The n semi-axes of the $(n - 1)$-dimensional isocontour ellipsoid formed by an n-dimensional Gaussian distribution are precisely the normalized eigenvectors of $\boldsymbol{\Sigma}$ multiplied by their eigenvalues.

If we let $\boldsymbol{\Sigma} = \mathbf{I}$, then this means that all the X_i's are pairwise uncorrelated since $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = 0$. In general, this does not mean that the X_i's are independent, but for joint Gaussians, this also implies independence!

Theorem 10.2. *Given multivariate Gaussian $\mathbf{X} = (X_1 \dots X_n)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the X_i's are pairwise independent if and only if they are uncorrelated.*

Proof. We can expand the pdf of $\mathbf{X}$ as

$$\begin{aligned}
f_X(x) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(x - \mu)^T(x - \mu) \right) \\
&= \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(\sum_{i=1}^n -\frac{1}{2}(x_i - \mu_i)^2 \right) \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu_i)^2 \right)
\end{aligned}$$

which is the product of n single-variable Gaussians X_i. Therefore this means that independence and uncorrelation are equivalent!   ∎

Therefore, if the nondiagonal entries of the covariance matrix are all 0, then we know that the variables are all uncorrelated and therefore independent.

11 Order Statistics

Let $X_1, X_2, ..., X_n$ be a finite collection of independent, identically distributed random variables. Suppose that they are continuously distributed with density f and CDF F.

Definition 11.1. Define the random variable $X_{(k)}$ to be the kth ranked value, called the *kth order statistic*. This means that

$$X_{(1)} = \min\{X_1, X_2, ..., X_n\}, \ \ X_{(n)} = \max\{X_1, X_2, ..., X_n\}$$

and in general, for any $k \in \{1, 2, ..., n\}$,

$$X_{(k)} = X_j \text{ if } \sum_{l=1}^{n} \mathbb{I}_{X_l < X_j} = k - 1$$

which means that exactly $k - 1$ of the values of X_l are less than X_j. Since F is continuous,

$$X_{(1)} < X_{(2)} < ... < X_{(n)}$$

holds with probability 1. This leads us to define the random variable $X_{(k)}$ representing the kth order statistic.

$$f_{(k)}(y) = \begin{cases} n \binom{n-1}{k-1} y^{k-1}(1-y)^{n-k} & y \in (0,1) \\ 0 & y \notin (0,1) \end{cases}$$

That is, $X_{(k)}$ has the Beta$(k, n - k_1)$ distribution.

11.1 Poisson Arrival Process

A *Poisson Arrival Process* with rate $\lambda > 0$ on the interval $[0, \infty)$ is a model for the occurence of some events which may have at any time. We can interpret the process as a collection of random points in $[0, \infty)$ which are the times at which the arrivals occur.

    Interpretation 1 Set $T_0 = 0$. The arrival times are random variables $0 < T_1 < T_2 < T_3 < ...$ such that the inter-arrival waiting times

$$W_k = T_k - T_{k-1}, \ \ \ k \geq 0$$

have the property that $\{W_k\}_{k=1}^{\infty}$ are independent Exp(λ) random variables.

Interpretation 2 For any interval $I \subset [0, \infty)$, let

$$N_I \equiv \text{ number of arrivals that occur in interval } I$$

Then, $N_I \sim$ Poisson$(\lambda|I|)$, and for any collection of disjoint intervals $I_1, I_2, ..., I_n$, the random variables

$$\{N_{I_k}\}_{k=1}^{n}$$

are independent.

Theorem 11.1. *These two interpretations of the arrival process are equivalent.*

Proof. In the 2nd interpretation, the statement $N_I \sim$ Poisson$(\lambda|I|)$ means that

$$\mathbb{P}(N_I = m) = e^{-\lambda|I|} \frac{(\lambda|I|)^m}{m!}, \ \ \ m = 0, 1, 2, 3, ...$$

where $|I|$ is the length of interval I. From the first perspective, notice that

$$T_k = W_1 + W_2 + ... + W_k$$

so that the kth arrival time T_k is a sum of k independent Exp(λ) random variables. Thus,

$$T_k \sim \text{Gamma}(k, \lambda)$$

and therefore has density

$$\lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!}, \ \ \ t > 0$$

Note that the arrival times T_i are not independent of each other, but the wait times W_i are indeed independent. ∎

We can slightly modify this to create a Poisson arrival process over some finite time horizon $[0, L]$. Again, you can do this two ways:

1. Starting with independent $\text{Exp}(\lambda)$ random variables $W_1, W_2, ...$, we define

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

   Once you have $T_k > L$, stop.

2. We let $N \sim \text{Poisson}(\lambda L)$, since we are only working in finite interval L. Given $N = n$, let $U_1, U_2, ..., U_n \sim \text{Uniform}([0, L])$. These define the arrival times, and let us order them to get

$$T_k = U_{(k)}, \quad k = 1, 2, ..., N$$

   where $U_{(k)}$ is the kth ordered point, with $T_1 = \min(U_1, ..., U_N)$.

Lemma 11.2 (Memoryless Property)**.** *The $\text{Exp}(\lambda)$ distribution has the property that for all $t, s \geq 0$,*

$$\mathbb{P}(W > t + s \mid W > t) = \mathbb{P}(W > s)$$

which is called the memoryless property. We can interpret this in the following way. Let W be the time you have to wait for the first arrival. Given that you already waited t units of time, the probability that you have the wait s additional units of time is just the probability that you wait at least s from the beginning. That is, knowing that t units of time have elapsed does not affect the distribution of the remaining waiting time.

Theorem 11.3. *Let W be a continuously distributed random variable. Then $W \sim \text{Exp}(\lambda)$ for some $\lambda > 0$ if and only if W satisfies the memoryless property.*

12 Markov Chains

12.1 Discrete Time Chains

Definition 12.1. A *Markov chain* is a sequence of random variables $\{X_n\}_{n=0}^{\infty}$, which take values in some set $\mathcal{S}$, called the *state space* satisfying the *Markov property*. Since we are working with discrete time chains, we will assume that $\mathbb{S}$ is a countable (and in most cases, finite). Thus, the X_n will all be discrete random variables. We can also think of X_n as a discrete "time" index; that is, X_n is the state of the system at time n. Therefore, the sequence of random variables models a system evolving in a random way.

Definition 12.2. A sequence of random variables $\{X_i\}$ satisfies the *Markov property* if

$$\mathbb{P}(X_{n+1} = y \mid X_n = x_n, X_{n-1} = x_{n-1}, ..., X_0 = x_0\} = \mathbb{P}(X_{n+1} = y \mid X_n = x_n\}$$

holds for any choice oc states $y, x_n, x_{n-1}, ..., x_0 \in \mathcal{S}$ and for any $n \geq 1$.

Colloquially, given that one is at state $X_n = x_n$, knowing all the previous states does not help in predicting X_{n+1}. Knowing only the current state is relevant in predicting the next one. We can model this entire system using a matrix.

Definition 12.3. Assuming that the chain is *time-homogeneous*, the *transition probability matrix* P has elements P_{xy} defined

$$P_{xy} = P(x, y) = \mathbb{P}(X_1 = y \mid X_0 = x) = \mathbb{P}(X_{n+1} = y \mid X_n = x)$$

which is the probability of moving from state x to state y in one step. The time homogeneous condition refers to the last equality; that is, the one-step transition probabilities don't change with the time index n. Note that if $\mathcal{S}$ is finite, then P is a $|S| \times |S|$ matrix, and if $\mathcal{S}$ is countably infinite, then P is an infinite-dimensional matrix. The axioms of probability imply that A^T is an entry-wise nonnegative stochastic matrix.

Example 12.1 (Random Walks)**.** *A random walk on the integers $\mathcal{S} = \mathbb{Z}$ where a point has equal probability of moving right or left can be modeled with the probability function.*

$$P(x, y) = \mathbb{P}(X_{n+1} = y \mid X_n = x) = \begin{cases} \frac{1}{2} & y = x + 1 \\ \frac{1}{2} & y = x - 1 \\ 0 & otherwise \end{cases}$$

This can be generalized to multiple dimensional random walks on graphs with probability function

$$P(x, y) = \frac{1}{deg(x)}$$

where $deg(x)$ is the number of adjacent nodes to node x. In this way, the point hops randomly from node to node, and if the graph is connected, then the walker can visit any vertex in the graph.

Example 12.2 (Discrete Moran Model)**.** *Consider a population of size N. Each individual is one of two types (say, red or blue). At each time step, the system evolves in the following way: First, one of the individuals is chosen uniformly at random to be eliminated from the population; and another individual is chosen uniformly at random to produce one offspring identical to itself. These two choices are made independently. So, if a red individual is chosen to reproduce, and a blue one is chosen for elimination, then the total number of red particles increases by one and the number of blue particles decreases by one. If a red is chosen for reproduction and a red is chosen for elimination, then there is no net change in the number of reds and blues. Let X_n be the number of red individuals at time n. The transition matrix for this chain is*

$$P_{ij} = \begin{cases} \frac{i}{N}\left(\frac{N-i}{N}\right) & j = i - 1, i \neq 0 \\ \left(\frac{N-i}{N}\right)\frac{i}{N} & j = i + 1, i \neq N \\ 1 - 2\left(\frac{N-i}{N}\right)\frac{i}{N} & j = i \\ 0 & otherwise \end{cases}$$

Note that the states $X_n = 0$ and $X_n = N$ are absorbing states, which represents a phenomenon called fixation.

Definition 12.4. A certain state F in the state space $\mathcal{S}$ of a Markov chain is called an *absorbing state* if

$$\mathbb{P}(X_{n+1} = F \mid X_n = F) = 1 \iff \mathbb{P}(X_{n+1} \neq F \mid X_n = F) = 0$$

Theorem 12.1. *Let there exist a time homogeneous Markov chain with transition probability matrix P. Given a probability distribution ν_n (a row vector) representing the a state of a system at time $t = n$, the probability distribution of which state the system will be at when $t = n + 1$ can be calculated by*

$$\nu_{n+1} = \nu_n P$$

The probability distribution of the state of the system at $t = n + k$ can be calculated by summing up all of the possible probabilities that lead to each state at $t = n + k$. It is calculated equivalently as matrix multiplication:

$$\nu_{n+k} = \nu_n P^k$$

Definition 12.5. The distribution ν of a Markov chain at time $t = 0$ is called the *initial distribution* for the chain. That is, ν is the initial distribution if

$$\mathbb{P}(X_0 = x) = \nu(x)$$

Definition 12.6. An *invariant distribution*, or *stationary distribution*, is a probability distribution π such that

$$\pi P = \pi$$

This means that

$$\pi P^k = \pi$$

for all $k \in \mathbb{N}$. We can equivalently call π the left eigenvector of matrix P with eigenvalue 1. If π is an invariant distribution for the chain, and $X_0 \sim \pi$, then the distribution of X_n does not change with n; it is invariant. Note that this does not mean that X_n is constant; rather, it means that the distribution of X_n is not changing.

Example 12.3. *Let us have a two node system with nodes labeled L and R. That is, $\mathcal{S} = \{L, R\}$. Consider a chain on this state space with transition probability matrix.*

$$P = \begin{pmatrix} 1 - a & a \\ b & 1 - b \end{pmatrix}$$

which can be visualized in the following diagram below.
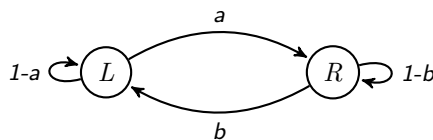
Then, the stationary distribution is

$$\pi = \left(\frac{b}{a + b}, \frac{a}{a + b} \right)$$

Notice that if $a = b = 0$, then this definition is ill-defined, and any probability distribution is invariant since $P = I_2$, the identity matrix.

Definition 12.7. A state $x \in \mathcal{S}$ is *recurrent* if

$$\mathbb{P}(X_n = n \text{ for some } n \geq 1 \mid X_0 = x) = 1$$

That is, if the initial state is x, the chain has probability 1 of returning to x at some later time. If a state is not recurrent, then the state is said to be *transient*. That is, if x is transient, there is some positive probability that the chain will never return to x.

Definition 12.8. Two states $x, y \in \mathcal{S}$ are said to *communicate*, denoted $x \leftrightarrow y$, if there are positive integers n and m such that

$$P^{(n)}(x, y) > 0 \text{ and } P^{(m)}(y, x) > 0$$

That is, there is some positive probability that the chain can go from x to y and from y to x in some number of steps.

Definition 12.9. If all pairs $x, y \in \mathcal{S}$ communicate, then the chain is said to be *irreducible*. If there exists a pair of states that do not communicate, then the chain is said to be *reducible*.

Note that the notion of communication is an equivalence relation between states. That is, it satisfies the properties.

1. $x \leftrightarrow x$.

2. $x \leftrightarrow y \implies y \leftrightarrow x$.

3. $x \leftrightarrow y, y \leftrightarrow z \implies x \leftrightarrow z$.

This relation partitions the state space $\mathcal{S}$ uniquely into transient states and irreducible sub-chains

$$\mathcal{S} = T \cup C_1 \cup C_2 \cup ...$$

More specifically, T is the set of all transient states, and the sets C_k are *closed communication classes*, meaning that

1. For all $x, y \in C_k$, $x \leftrightarrow y$.

2. $P(x, z) = 0$ whenever $x \in C_k$ but $z \notin C_k$.

Note that for all $x, y \notin T$, x and y communicate if and only if x and y are in the same class C_k. Moreover, once the chain reaches one of the sets C_k, it cannot leave C_k.

Definition 12.10. For any state $x \in \mathcal{S}$, the *period* of x is defined to be

$$d(x) \equiv \gcd\{n \geq 1 \mid P^{(n)}(x, x) > 0\}$$

Theorem 12.2. *It follows that if two states x and y communicate, then they must have the same period: $d(x) = d(y)$. It naturally follows that if the chain is irreducible, then all states must have the same period, and we can define the period of the chain to be $d(x)$ for any x we choose.*

Definition 12.11. If an irreducible chain has period 1, the chain is said to be *aperiodic*. Otherwise, the chain is *periodic* with period $d > 1$.

Theorem 12.3. *Suppose $|\mathcal{S}| < \infty$. If the chain is irreducible, then there always exists a unique stationary distribution π. If the chain is also aperiodic, then for any initial distribution ν,*

$$\lim_{k \to \infty} \nu P^k = \pi$$

Hence

$$\lim_{k \to \infty} P^{(k)}(x, y) = \pi(y)$$

for all $x, y \in \mathcal{S}$. Furthermore, for any function $F : \mathcal{S} \longrightarrow \mathbb{R}$, the limit

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} F(X_n) = \sum_{x \in \mathcal{S}} F(x)\, \pi(x) = \mathbb{E}\big(F(x)\big)$$

holds with probability 1. In particular, the limit does not depend on the initial distribution.

Proof. The Frobenius Extension to Perron's theorem (Linear Algebra, Theorem 7.31) combined with its applications to stochastic matrices (Linear Algebra, Theorem 7.30) proves this statement. ∎

Definition 12.12. For each $x \in \mathcal{S}$, define the *first visit* to x by

$$T_x \equiv \min\{n \geq 1 \mid X_n = x\}$$

This T_x is an integer-valued random variable. We say $T_x = +\infty$ if X_n never reaches x. Then, we define the *mean return time* to x by

$$\mu_x \equiv \mathbb{E}\big(T_x \mid X_0 = x\big)$$

If x is transient, then $\mu_x = +\infty$, since there is positive probability that $T_x = +\infty$.

Definition 12.13. It is possible that x is recurrent while $\mu_x = +\infty$. If this is the case, then x is said to be *null-recurrent*. If x is recurrent and $\mu_x < \infty$, then x is said to be *positive recurrent*.

Theorem 12.4. *An irreducible chain has a stationary probability distribution π if and only if all states are positive recurrent. If a chain is irreducible and all states are positive recurrent, then*

$$\pi(x) = \frac{1}{\mu_x}$$

for all $x \in \mathcal{S}$. π is also unique.

12.1.1　Exit Probabilities

Suppose a chain is finite and irreducible. Let $a, b \in \mathcal{S}$ be given states, and let us define $h(x)$ to be the probability of hitting b before a, given that we start from x.

$$h(x) \equiv \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_0 = x)$$

Clearly, $h(b) = 1$ and $h(a) = 0$. By conditioning on the first jump out of x, we also have

$$\begin{aligned}
h(x) &= \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_0 = x) \\
&= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_1 = y, X_0 = x)\, \mathbb{P}(X_1 = y \mid X_0 = x) \\
&= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_1 = y, X_0 = x)\, P(x, y) \\
&= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_1 = y)\, P(x, y) \\
&= \sum_y h(y)\, P(x, y)
\end{aligned}$$

The sum is over all $y \in \mathcal{S}$ for which $P(x, y) \neq 0$. This gives us a linear system of equations to solve for h

$$h(x) = \sum_y P(x, y) h(y) \ \forall x \in \mathcal{S} \setminus \{a, b\},$$
$$h(b) = 1,$$
$$h(a) = 0$$

12.1.2 Exit Prize

Let $B \subset \mathcal{S}$ be some subset of the state space, and let $g : B \longrightarrow \mathbb{R}$ be some function. Consider the function
$$h(x) = \mathbb{E}\big(g(X_\tau) \,|\, X_0 = x\big)$$
where $\tau = \min\{n \geq 0 \,|\, X_n \in B\}$ is the first time that the chain reaches some state in the set B (this time is random). We can interpret $g(y)$ as a "prize" that is awarded if the chain first reaches B at state y, which means that $h(x)$ is the expected prize, given that $X_0 = x$. If $x \in B$, then $\tau = 0 \implies h(x) = g(x)$. But if $x \notin B$, then by the same argument as shown in exit probabilities, it is true that h satisfies the linear system of equations

$$h(x) = \sum_g P(x, y) h(y), \quad \forall x \in \mathcal{S} \setminus B,$$
$$h(x) = g(x), \quad x \in B$$

Note that Exit probability system is a special case of the Exit prize system. In the former, we have defined $B = \{a, b\}$ and g defined by $g(a) = 0, g(b) = 1$.

12.1.3 Occupation Times, Absorbing States

Suppose that a chain on a finite $\mathcal{S}$ is irreducible. Let $B \subset \mathcal{S}$ be some subset of states and let $A = \mathcal{S} \setminus B$ be the other states. Then for $x \in A$, we wish to know how many steps the chain will take before reaching a state in the set B. We define
$$\tau_B = \min\{n \geq 0 \,|\, X_n \in B\}$$
which represents the first time that X is in B, an integer valued random variable. We wish to compute

$$h(x) = \mathbb{E}(\tau_B \,|\, X_0 = x)$$

Clearly, $h(y) = 0$ for all $y \in B$. For $x \in A$, it takes at least one step to reach $B \implies h(x) \geq 1$ for $x \in A$. We condition on the first step from x. This leads to the system

$$h(x) = 1 + \sum_{y \in \mathcal{S}} P(x, y) \, \mathbb{E}(\tau_B \,|\, X_1 = y), \forall x \in A = \mathcal{S} \setminus B$$

Since the chain is time-homogeneous, this means that

$$h(x) = 1 + \sum_{y \in \mathcal{S}} P(x, y) \, h(y), \forall x \in A$$

Since $h(y) = 0$ for all $y \in B$, we now have

$$h(x) = 1 + \sum_{y \in A} P(x, y) \, h(y), \forall x \in A$$

To solve this system, let us define M as the $|A| \times |A|$ submatrix of P obtained by keeping only the entries $P(x, y)$ with $x, y \in A$. So, the system can be written as

$$h(x) = 1 + \sum_{y \in A} M(x, y) \, h(y), \forall x \in A$$

We can solve this system of equations through the equivalent matrix equation

$$(I - M)h = 1$$

where $1 = (1, 1, ..., 1)^T$ is the column vector consisting of all 1's. The solution vector is therefore

$$h = (I - M)^{-1}1$$

So, for a particular $x \in A$,

$$h(x) = \sum_{y \in A} (I - M)^{-1}(x, y)$$

Alternatively, we can slightly modify the chain to chain $\tilde{X}_n$ by replacing the transition probability matrix P with another one defined as

$$\tilde{P}(x, y) = \begin{cases} P(x, y) & x \in A, y \in \mathcal{S} \\ 1 & x = y \in B \\ 0 & \text{else} \end{cases}$$

This modification means that all transitions from state in A to any other state are preserved and the only transitions from a state $x \in B$ are self loops. In particular, all transitions from states $x \in B$ to states $y \in A$ are removed. Therefore, under this modified transition matrix, the states in B are absorbing states. The tail sum formula implies that

$$\mathbb{E}(\tau_B \mid X_0 = x) = \sum_{k=0}^{\infty} \mathbb{P}(\tau_B > k \mid X_0 = x)$$

Notice that since the chain X_n and $\tilde{X}_n$ have the same transition rules before hitting a state B, we have

$$P^{(k)}(x, y) = \tilde{P}^{(k)} = M^{(k)}(x, y)$$

where M is the $|A| \times |A|$ submatrix defined previously. Therefore, putting this all together, we have

$$\begin{aligned} \mathbb{E}(\tau_B \mid X_0 = x) &= \sum_{k=0}^{\infty} \mathbb{P}(\tau_B > k \mid X_0 = x) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(\tilde{X}_k \in A \mid X_0 = x) \\ &= \sum_{k=0}^{\infty} \sum_{y \in A} \tilde{P}^{(k)}(x, y) \\ &= \sum_{k=0}^{\infty} \sum_{y \in A} M^{(k)}(x, y) \\ &= \sum_{y \in A} \left(\sum_{k=0}^{\infty} M^{(k)} \right)(x, y) \end{aligned}$$

Using a theorem from linear algebra, we can show that if all the eigenvalues of a $d \times d$ matrix M have modulus strictly less than 1, then $I - M$ is invertible and

$$\sum_{k=0}^{\infty} M^{(k)} = (I - M)^{-1}$$

where I is the $d \times d$ identity matrix. If M is the $|A| \times |A|$ submatrix described above, one can show that M has his property and that $I - M$ is invertible. Hence,

$$\mathbb{E}(\tau_B \mid X_0 = x) = \sum_{y \in A} \left(\sum_{k=0}^{\infty} M^{(k)} \right)(x, y) = \sum_{y \in A} (I - M)^{-1}(x, y)$$

which refers to the (x, y) entry of the matrix $(I - M)^{-1}$. This is indeed consistent with our previous derivation of the formula for $h(x)$, the expected number of steps before the state reaches B.

12.2   Markov Chain Monte Carlo Algorithms

In statistics, Markov chain Monte Carlo (MCMC) methods comprise of a class of algorithms for sampling from a probability distribution by constructing a Markov chain that has the desired distribution as its equilibrium distribution. That way, by recording samples from the chain, one may get better approximations of the actual distribution.

Let there exist a state space $\mathcal{S}$ with some probability distribution $\pi(x)$ for every $x \in \mathcal{S}$. Clearly,

$$\sum_{x \in \mathcal{S}} \pi(x) = 1$$

but the problem is that we do not know that π is. We do know, however, another function f that is directly proportional to π.

$$\pi(x) = \frac{f(x)}{c}, \text{ where } c = \sum_{x \in \mathcal{S}} f(x)$$

is the normalizing constant. It is often the case that c is unknown and the state space $\mathcal{S}$ is so large that computing c directly is expensive. Therefore, we construct Markov chains that can provide approximations to π.

12.2.1   Metropolis-Hastings Algorithm

This algorithm is useful because it does not require knowledge of the normalizing constant c. The algorithm only requires evaluations of

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

We first have the state space $\mathcal{S}$ consisting of all the possible states. We now construct (any) probability transition matrix q for a Markov chain on $\mathcal{S}$. Note that q is a $|\mathcal{S}| \times |\mathcal{S}|$ matrix and q^T is a stochastic matrix. This matrix is constructed by the user and is completely well-defined and known. We start off with any initial state $x_0 \in \mathcal{S}$ and iterate the following 2-steps to construct a Markov chain.

1. Given a state $X_n = x$, we generate a new state X_{n+1} by first proposing a new state $y \in \mathcal{S}$ with probability $q(x, y)$ (determined from the matrix q).

2. With this chosen state y, we decide whether to accept to reject the proposal. With probability

$$\min\left(1, \frac{\pi(y) \, q(y, x)}{\pi(x) \, q(x, y)}\right)$$

we accept the proposal and set $X_{n+1} = y$. Otherwise, the proposal is rejected and the new state is the same $X_{n+1} = x$.

Note that there are two levels of randomness here: which state the new state y will be and whether to accept this state to be the next one or not. If step two did not exist (i.e. the probability of accepting the proposal is always 1), then this would just be a regular Markov chain represented by the matrix q. But the addition of step 2 means that while q is used in constructing the discrete chain X_n, it is *not* the transition probability matrix of X_n.

There is also a lot of flexibility on choosing q, although the performance of the algorithm (speed of convergence of the distribution of X_n to the stationary distribution) will depend on the choice.

Proposition 12.5. *For the chain defined by the Metropolis-Hastings algorithm, the distribution π is stationary.*

Proof. Let us write in shorthand

$$\alpha(x, y) = \frac{\pi(y) \, q(y, x)}{\pi(x) \, q(x, y)}$$

First, observe that if $x \neq y$, the transition probability for the chain defined by the algorithm is just

$$P(x, y) = q(x, y) \, \min\{1, \alpha(x, y)\}$$

Next, we claim that for all $x, y \in \mathcal{S}$,

$$\pi(x)P(x, y) = \pi(y) \, P(y, x)$$

This condition is called *detailed balance*. Assuming that $\alpha(x, y) \leq 1$, it is true that

$$\pi(x)P(x, y) = \pi(x)q(x, y)\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = \pi(y)q(y, x)$$

In this case, we also have $\alpha(y, x) = 1/\alpha(x, y) \leq 1$. So,

$$\pi(y)P(y, x) = \pi(y)q(y, x)$$

and we have proved what we had claimed. Now, summing over x,

$$\sum_x \pi(x)P(x, y) = \sum_x \pi(y)P(y, x) = \pi(y)\sum_x P(y, x) = \pi(y)$$

since P^T is stochastic. ∎

12.2.2 Gibb's Sampling

Let $\mathcal{A} = \{a_1, ..., a_k\}$ be some finite set. Suppose that the state space

$$\mathcal{S} = \mathcal{A} \times ... \times \mathcal{A} = \mathcal{A}^M$$

for some $M \in \mathbb{N}$. The following algorithm generates a Markov chain on $\mathcal{S}$ with stationary distribution

$$\pi(x) = \frac{f(x_1, x_2, ..., x_M)}{c}, \quad x = (x_1, x_2, ..., x_M) \in \mathcal{S}$$

where $c > 0$ is a normalizing constant. Note that $|\mathcal{S}| = k^M$, so computing c may be expensive when M is large. The current state of the chain is denoted

$$X_n = (X_n^1, X_n^2, ..., X_n^M)$$

We think of X_n as having M components, each component taking values in $\mathcal{A}$. We start off with any initial state $X_0 = (X_0^1, X_0^2, ..., X_0^M)$ and construct a Markov chain by iterating the following two steps.

1. Given $X_n = (X_n^1, X_n^2, ..., X_n^M)$, we generate the next state X_{n+1} by picking a component index $i \in \{1, ..., M\}$ uniformly at random.

2. With this chosen, well-defined i, we choose a random $Y^i \in \mathcal{A}$ according to the distribution

$$\mathbb{P}(Y^i = a) = \frac{f\left(X_n^1, ..., X_n^{i-1}, a, X_n^{i+1}, ..., X_n^M\right)}{\sum_{j=1}^k f\left(X_n^1, ..., X_n^{i-1}, a_j, X_n^{i+1}, ..., X_n^M\right)}, \quad a \in \{a_1, ..., a_k\}$$

3. Then, set $X_{n+1} = \left(X_n^1, ..., X_n^{i-1}, Y^i, X_n^{i+1}, ..., X_n^M\right)$.

Note that at each step, only one component of X_n is updated. Observe that the distribution above is also equal to

$$\mathbb{P}(Y^i = a) = \frac{\pi\left(X_n^1, ..., X_n^{i-1}, a, X_n^{i+1}, ..., X_n^M\right)}{\sum_{j=1}^k \pi\left(X_n^1, ..., X_n^{i-1}, a_j, X_n^{i+1}, ..., X_n^M\right)}$$

which is the marginal distribution of the ith component, given the values of the other components.

Proposition 12.6. *For the chain defined by this algorithm, the distribution π is stationary.*

Proof. We verify that the detailed balance condition holds. It is also helpful to note that $P(x, y) \neq 0$ if and only if x and y differ in one coordinate. ∎

12.3 Continuous Time Markov Chains

As the name suggests, in a continuous time Markov chain X_t, the time parameter is continuous ($t \geq 0$). As before, the system jumps randomly between states in $\mathcal{S}$, but now the jumps may occur at any time and they occur randomly. This implies that there are *two* sources of randomness:

1. *where* the system jumps and

2. *when* the system jumps

Definition 12.14. The Markov property in the continuous time case says that for any $s, t \geq 0$ and $y \in \mathcal{S}$,
$$\mathbb{P}(X_{t+s} = y \,|\, X_t) = \mathbb{P}(X_{t+s} = y \,|\, X_r \ \forall 0 \leq r \leq t)$$

Colloquially, the conditional distribution of X_{t+s} given the history up to time t is the same as the conditional distribution of X_{t+s} given only X_t. Thus, if we know the current state at t, knowing information about the past doesn't help us better predict the future state X_{t+s}.

In order for the Markov property to hold, the times between jumps must be exponentially distributed random variables because it is the only density that has the memoryless property. This fact has already been stated in a theorem when covering Poisson arrival processes. This is what makes $\text{Exp}(\lambda)$ so important for continuous time Markov chains.

Lemma 12.7. *Let $T_1, T_2, ..., T_n$ be independent exponential random variables with rates $\lambda_1, \lambda_2, ..., \lambda_n$, respectively. Then the random variable $T \equiv \min\{T_1, T_2, ..., T_n\}$ is*

$$T \sim \text{Exp}\Big(\sum_{i=1}^{n} T_i \Big)$$

Moreover,

$$\mathbb{P}(T_k = \min\{T_1, ..., T_n\}) = \frac{\lambda_k}{\lambda_1 + ... + \lambda_n}$$

We can interpret the lemma above by imagining that we have n alarm clocks all set simultaneously, which will ring independently at random times. Suppose that clock k will ring after T_k units of time have expired, where T_k is a random variable distributed as $\text{Exp}(\lambda_k)$. Then, $T = \min\{T_1, ..., T_n\}$ is the time at which the first ring occurs.

Example 12.4. *The simplest and the most important continuous time Markov chains is the Poisson arrival process. The process really has a single parameter $\lambda > 0$ (the rate of process) by definition and is integer valued. At each jump time, the process increases by 1, and the time between jumps are independent, distributed as $\text{Exp}(\lambda)$.*

Notice that when λ is large, the arrivals occur more frequently than when λ is small, because the expected time between arrivals is $1/\lambda$. The second way we can interpret it is to choose an interval of time t and let X_t be the number of jumps that have occurred up to time t. It is a fact that X_t is a integer-valued, $\text{Poisson}(\lambda t)$ distribution. That is,

$$\mathbb{P}(X_t = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \ k = 0, 1, 2, ...$$

In particular, $\mathbb{E}(X_t) = \lambda t$ and $\text{Var}(X_t) = \lambda t$.

12.4 Branching Processes

Definition 12.15. A *branching process* is a type of Markov chain modeling a population in which each individual produces a random number of children (possibly 0) and dies. The state space is $\mathcal{S} = \{0, 1, 2, 3, ...\}$. Furthermore, there is a discrete-time version and a continuous time version of the chain. In the discrete case, the state is Z_n, the size of the population at time $n = 0, 1, 2, ...$, and in the continuous case, the state is Z_t for $t \geq 0$.

### 12.4.1 Discrete-time Branching Process

In the discrete case, all of the Z_n individuals in the current generation branch at the same time and immediately die. The branching is independent and distributed according to the *offspring distribution* $\{p_k\}_{k=0}^\infty$. Specifically, if $Z_n = m$, then

$$Z_{n+1} = Y_1^n + Y_2^n + ... + Y_m^n$$

where Y_i^n represents the number of offspring the ith individual in the nth generation has. All of them are distributed as

$$\mathbb{P}(Y_i^n = k) = p_k, \ k = 0, 1, 2, 3, ...$$

where p_k is the probability that a parent has k children. Note that if $p_0 \neq 0$, then there is positive probability that $Y_i^n = 0$ for all i, meaning that the population can go extinct. A sample branching process up to the second generation is shown below.
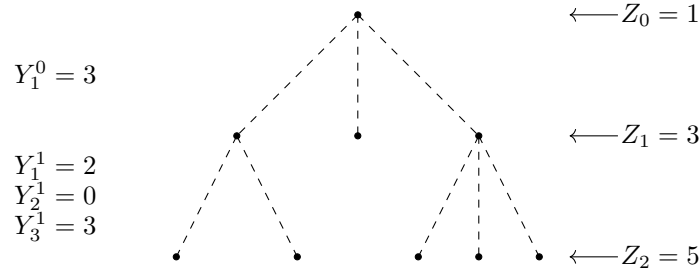
Suppose that the mean number of offspring of a single parent is finite.

$$\mu = \mathbb{E}(Y) = \sum_{k=0}^\infty k \, \mathbb{P}(Y = k) = \sum_{k=0}^\infty k \, p_k < \infty$$

If Y_1 and Y_2 are two independent, discrete random variables, we can define their convolution and use the fact that $\mathbb{P}(Y_i = k) = p_k$ to get

$$\mathbb{P}(Y_1 + Y_2 = k) = \sum_j \mathbb{P}(Y_1 = k - j) \, \mathbb{P}(Y_2 = j)$$

$$= \sum_{j=0}^\infty p_{k-j} p_j, \ \ k = 0, 1, 2, ...$$

This is a two-fold convolution of the sequence $\{p_k\}$ with itself, denoted

$$p_k^{*2} = \sum_{j=0}^\infty p_{k-j} \, p_j$$

Extending this, we can find the m-fold convolution of the sequence $\{p_j\}$ with itself, represented by the sequence $\{p_j^{*m}\}$, where p_k^{*m} is the kth term in this sequence. This gives us

$$p_k^{*n+1} = \sum_{j=0}^\infty p_{k-j} \, p_j^{*n}$$

for all $n \in \mathbb{N}$. Using this, we can write down the transition probabilities for the Markov chain Z_n.

$$\mathbb{P}(Z_{n+1} = k \mid Z_n = m) = \begin{cases} 0 & \text{if } m = 0 \\ p_k^{*m} & \text{if } m \geq 1, k \geq 0 \end{cases}$$

where $\mathbb{P}(Z_{n+1} = k \mid Z_n = m)$ represents the probability of the nth generation consisting of m individuals producing a total of k offspring for the $(n+1)$th generation. Thus, the branching process is completely determined by the distribution of Z_0 and the offspring distribution $\{p_k\}_{k=0}^\infty$.

Lemma 12.8. *Given this discrete-time branching process, let μ be the mean of the offspring distribution. Then,*

$$\mathbb{E}(Z_n \mid Z_0 = 1) = \mu^n$$

If $\mu > 1$, the mean of Z_n grows exponentially, and if μ_1, the mean of Z_n decreases exponentially.

12.4.2 Continuous-time Branching Process

A continuous time branching process Z_t has very similar structure to the discrete time branching process, except that the times between branch events (for each individual) are independent exponentially distributed random variables $\text{Exp}(\lambda)$, where the parameter $\lambda > 0$ is the branching rate. It is as though each individual has an independent alarm clock which rings as a time that is $\text{Exp}(\lambda)$, independently of all other clocks. So, if there are currently N individuals, then the next alarm will ring at rate λN; that is, the time until the next ring is distributed as $\text{Exp}(\lambda N)$, since it is the minimum of N independent $\text{Exp}(\lambda)$ random variables. When an individual branches (clock rings), that individual produces a random number of offspring, according to the offspring distribution $\{p_k\}$, as before. So, a continuous time branching process has the same geneological structure as the discrete time process, but the times between branch events is randomized. Consequently, whether or not the process eventually goes extinct, depends only on the offspring distribution, not on the branching rate λ.

Let $m_1(t) = \mathbb{E}(Z_t)$ denote the expected population size at time t. Then, it is a fact that $m_1(t)$ satisfies the ordinary differential equation

$$\frac{d}{dt}m_1(t) = \lambda(\mu - 1)m_1(t)$$

where

$$\mu = \sum_{k=1}^{\infty} k p_k$$

is the mean of the offspring distribution. Solving this equation reveals that

$$m_1(t) = e^{\lambda(\mu-1)t}m_1(0)$$

If $\mu > 1$, the mean population size grows exponentially, and if $\mu < 1$, the mean population size decreases exponentially.

12.4.3 Extinction Probability, Generating Functions

The expression for the transition probabilities of Z_n (disrete case) is quite difficult to work with. Alternatively, it can be convenient to work with generating functions.

Definition 12.16. The *generating function* for the offspring distribution is the function

$$G(s) \equiv \sum_{k=0}^{\infty} p_k\, s^k = \mathbb{E}(s^Y)$$

where $Y \sim \{p_k\}$ is a random variable representing the number of children produced by a given individual. Note that G is a power series that simply encodes information about the offspring distribution (also a sequence) $\{p_k\}_{k=0}^{\infty}$.

Theorem 12.9.
  1. *The radius of convergence of $G(s)$ is at least 1. $G(s)$ defines a continuous function on $|s| \leq 1$.*

  2. *On the interval $[0,1]$, $G(s)$ is increasing and convex. If $p_0 + p_1 < 1$, then $G(s)$ is strictly convex for $s \in [0,1]$.*

  3. *$G(0) = p_0$.*

  4. *$G(1) = 1$.*

  5. *$G'(1^-) = \mu$ is the expected number of offspring of a single individual.*

Proof. We use the fact that

$$\sum_{k=0}^{\infty} p_k = 1 \text{ and } 0 \geq p_k \geq \quad \forall k = 0, 1, 2, \ldots$$

∎

Theorem 12.10. *Suppose that $Z_0 = 1$ and that $p_0 + p_1 < 1$. Then*

$$\lim_{n \to \infty} \mathbb{P}(Z_n = 0) = \mathbb{P}(\text{eventual extinction}) = t$$

where $t \in [0, 1]$ is the smallest non-negative root of the equation $t = G(t)$. If $\mu \leq 1$, then $t = 1$ (clearly, since the population will exponentially decrease on average). If $\mu > 1$, there is a positive probability that the population never goes extinct.

Proof. Let t be the probability that an individual's descendent family tree goes extinct. That is, $t = \mathbb{P}(Z_n = 0 \text{ for some } n \geq 1 \mid Z_0 = 1)$. To derive the equation $t = G(t)$, let us condition on the first generation, with Y_1 denoting the number of offspring of the single parent.

$$t = \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1)$$

$$= \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k)\, \mathbb{P}(Y_1 = k \mid Z_0 = 1)$$

$$= \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k)\, p_k$$

That is, given that there are k children of the first individual, the probability that this first individual's descendent family tree will go extinct is equal to the probability that each of the k children's trees go extinct. These k extinction events are independent. Therefore,

$$\mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k) = t^k$$

which implies that

$$t = \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k)\, p_k = \sum_{k=0}^{\infty} t^k\, p_k = G(t)$$

Additionally, under the hypothesis that $p_0 + p_1 < 1$, then $G(s)$ is strictly convex on $[0, 1]$. Hence if $G'(1) = \mu \leq 1$, the smallest non-negative root of $t = G(t)$ must be $t = 1 \implies$ extinction occurs with probability 1. On the other hand, if $G'(1) = \mu > 1$, then the smallest root of $t = G(t)$ occurs in the interval $[0, 1)$. ∎

Note that this result applies to both the discrete time case and the continuous time case. In continuous-time chains, whether or not the population goes extinct does not depend on λ, the rate at which individuals give birth. The λ affects the time at which extinction occurs (if it occurs), but it does not affect the probability that it occurs. However, the extinction probability certainly does depend on the offspring distribution.

Definition 12.17. A random variable X is a *counting variable* if it takes values in $\{0, 1, 2, ...\}$.

Note that generating functions is a mapping from X, the set of counting variables (all assumed to be pairwise independent) to the algebra of power series over variable s.

$$G : X \longrightarrow F[[s]]$$

Lemma 12.11. *Let X and Y be two independent random counting variables, with generating functions $G_X(s) = \mathbb{E}(s^X)$ and $G_Y(s) = \mathbb{E}(s^Y)$. Then, the generating function for the random variable $Z = X + Y$ is $G_Z(s) = G_X(s)G_Y(s)$. That is, the generating function mapping G is a homomorphism that maps addition to multiplication. In particular, if X and Y are iid, then $G_Z(s) = G_X(s)^2$.*

Proof. Since X and Y are independent,

$$G_Z(s) = \mathbb{E}(s^Z) = \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X s^Y) = \mathbb{E}(s^X)\mathbb{E}(s^Y) = G_X(s)G_Y(s)$$

∎

Applying this argument iteratively, we get the following lemma.

Lemma 12.12. *Let $N \geq 1$ be a fixed positive integer. Let $Y_1, Y_2, ..., Y_N$ be independent, identically distributed random counting variables with generating function $G_Y(s) = \mathbb{E}(s^Y)$. Then, the generating function for the sum $Z = Y_1 + ... + Y_n$ is*

$$G_Z(s) = G_Y(s)^N$$

Now, suppose that N is not fixed, but another random variable. We wish to describe the distribution of the sum of a random number of random variables.

Lemma 12.13. *Let $Y_1, Y_2, Y_3, ...$ be a collection of independent, identically distributed random variables with generating function $G_Y(s) = \mathbb{E}(s^Y)$. Let N be a random counting variable, independent of the Y_i. Let N have generating function $G_N(s)$. Then the generating function for $Z = Y_1 + Y_2 + ... + Y_N$ is*

$$G_Z(s) = G_N\big(G_Y(s)\big)$$

Proof. Just condition on $N = k$

$$
\begin{aligned}
G_Z(s) = \mathbb{E}(s^Z) &= \sum_{k=0}^{\infty} \mathbb{E}\big(s^Z \mid N = k\big) \, \mathbb{P}(N = k) \\
&= \sum_{k=0}^{\infty} \mathbb{E}(s^{Y_1 + ... + Y_k} \mid N = k) \, \mathbb{P}(N = k) \\
&= \sum_{k=0}^{\infty} G_Y(s)^k \, \mathbb{P}(N = k) \\
&= \mathbb{E}\big(G_Y(s)^N\big) = G_N\big(G_Y(s)\big)
\end{aligned}
$$

$\blacksquare$

Theorem 12.14. *Let $G(s)$ be the generating function for the offspring distribution $G(s) = \sum_{k=0}^{\infty} p_k s^k$. Suppose that $Z_0 = 1$ and let $G_n(s) = \mathbb{E}(s^{Z_n})$ be the generating function for the random variable Z_n. Then,*

$$G_{n+m}(s) = G_n\big(G_m(s)\big) = G_m\big(G_n(s)\big)$$

Hence,

$$G_n(s) = G(G(G(...(G(s))...)))   \quad \text{n-fold composition}$$

Example 12.5. *Suppose the offspring distribution is*

$$p_k = qp^k, \ \ k \geq 0$$

for some $p \in (0,1)$, where $q = 1 - p$. Thus, the number of children from a given parent is $Y = X - 1$, where $X \sim Geom(q)$. Then, $\mathbb{E}(Y) = \frac{1}{q} - 1 = \frac{p}{q}$. With some computation, this means that

$$G(s) = \frac{q}{1 - ps}$$

and $t = \min\{1, \frac{q}{p}\}$.