Probability and Stochastic Processes Notes (ORF 309)

Max Chien

February 2024

Contents

Introduction

This document contains notes taken for the class ORF 309: Probability and Stochastic Processes at Princeton University, taken in the Spring 2024 semester. These notes are primarily based on lectures and lecture notes by Professor Mark Cerenzia. Other references used in these notes include the 2016 lecture notes by Professor Ramon van Handel, *Introduction to Probability Models* by Sheldon Ross, *Fundamentals of Probability* by Saeed Ghahramani, *Markov Chains* by J.R. Norris, and *A First Course in Stochastic Processes* by Samuel Karlin and Howard Taylor. Since these notes were primarily taken live, they may contains typos or errors.

Chapter 1

Basic Probability Theory

1.1 Introduction

Probability theory deals with quantifying *random* or *nondeterministic* events; that is, experiments where the outcome is unknown prior to conducting the experiment. We will approach this quantification from an axiomatic approach, in order to codify what we mean by the *probability* of an event. Historically, there have been two major theories of probability:

- (Objectivist) The probability of an event is the limit of the relative frequency of the event when repeating an event infinitely (frequentist approach).
- (Subjectivist) The probability of an event is the degree of confidence which we lend to the event happening (Bayesian approach).

However, both of these approaches are still nonrigorous and can lead to incorrect results.

Example 1.1

A fair coin is flipped twice. What is the probability of at least 1 heads?

Suppose we take the frequentist approach. Then we might define the probability $\mathbb{P}(E) \coloneqq$ (# of outcomes where E occurs)/(# of outcomes). But applying this definition to our problem, we might intuit that there are three possible numbers of heads flipped: that is, 0, 1, or 2. Then we might incorrectly declare the probability of at least 1 heads to be 2/3.

Thus, we are motivated to develop an axiomatic approach to probability that will allow us to unambiguously develop a theory of probability, and allow us to identify when our intuition is and is not correct.

1.2 Axioms of Probability

To begin developing theory that allows us to define the probability of an event, we must first define what we mean by an event.

Definition 1.1

A **random experiment** R is an experiment where all possible outcomes are known ahead of time, but which outcome actually occurs is unknown. The **sample space** Ω_R of an experiment R is the set of all possible outcomes. An **event** is a subset of the possible outcomes $E \subseteq \Omega$. If, after conducting the experiment, an outcome ω occurs, then we say E occurs if and only if $\omega \in E$.

It is important to note here that "all possible outcomes" is specifically distinct from "observable outcomes," that is, there may be something about the experiment which prevents us from differentiating multiple outcomes or observing if an outcome has occurred. This does not change our definition here. In order to quantify the notion of *observable* outcomes, we need a new structure:

Definition 1.2

A **σ-algebra** $\mathcal{F}$ is a collection of subsets of Ω satisfying:

- $\Omega \in \mathcal{F}$
- $A \in \mathcal{F} \implies A^c \in \mathcal{F}$ (the **complement** of A is $A^c \coloneqq \Omega \setminus A$)
- $A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_0^\infty A_i \in \mathcal{F}$

In other words, a sigma algebra contains the set of all possible outcomes and is closed under complements and countable unions.

The last notion that we need to formalize is the probability of an event. Rather than construct a specific rule for the probability of an event, we once again axiomatize our definition:

Definition 1.3

A **probability rule** or **measure** on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that

- $0 \leq \mathbb{P}(E) \leq 1$ for $E \in \mathcal{F}$ (this is redundant but important)
- $\mathbb{P}(\Omega) = 1$
- If $A, B \in \mathcal{F}$ are disjoint, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

For any event $E \in \mathcal{F}$, $\mathbb{P}(E)$ denotes the **probability** of E.

At this point we have formalized all the notions that we need to completely define random events and probabilities of those events. We can combine them into a single space:

Definition 1.4

A triple $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-algebra on Ω and $\mathbb{P}$ is a probability rule on $(\Omega, \mathcal{F})$ is called a **probability space**.

Example 1.2

Let Ω be a finite sample space and suppose all outcomes are observable and equally likely. Then we have $\mathcal{F} = \{\text{all subsets of } \Omega\} = 2^{\Omega}$ and $\mathbb{P}(E) \coloneqq |E|/|\Omega|$.

Note that although the collection of all subsets of Ω is indeed a σ-algebra on Ω, a σ-algebra need not consist of all possible subsets.

Definition 1.5

Let $\mathcal{A}$ be a collection of subsets of Ω. Then $\sigma(\mathcal{A})$ denotes the smallest σ-algebra containing $\mathcal{A}$; in other words, it is the σ-algebra generated by $\mathcal{A}$.

For instance, suppose we have a countable partition of Ω, $B_1, B_2, \ldots$. Then $\mathcal{G} = \sigma(B_1, B_2, \ldots) = \{\bigcup_{m \in I} B_i\}_{I \subseteq \mathbb{N}}$ is the collection of all possible unions of the B_i (including the empty union).

The following are basic properties of probability rules as formalized above.

Theorem 1.1

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the following hold:

- For any $A \in \Omega$, $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$.
- If $A \subseteq B$, then $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$, so $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

1.3 Conditional Probability

Given two events, $A, B \in \mathcal{F}$, we have previously defined the probability of each event independently: $P(A), P(B)$. But we may also be interested in how the two probabilities are related. We introduce the notion of conditional probability:

Definition 1.6

Given two events $A, B \in \mathcal{F}$, with $\mathbb{P}(B) > 0$, the **conditional probability** of A given B is $\mathbb{P}(A|B) \coloneqq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$. In particular, we can define an associated probability rule $\mathbb{P}_B(A) \coloneqq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$. In this case, $(\Omega, \mathcal{F}, \mathbb{P}_B)$ is a probability space.

By rearranging the above formula, we can see that $\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B)$. This fact can be extended to further events via the law of multiplication:

Theorem 1.2: Law of Multiplication

Given n events $A_1, A_2, \ldots, A_n$, with $\mathbb{P}(A_i) > 0$ for $1 \leq i \leq n$, the following law holds:

$$\mathbb{P}(\bigcap_{i=1}^{n} A_i) = \mathbb{P}(A_1) * \mathbb{P}(A_2|A_1) * \mathbb{P}(A_3|A_1 \cap A_2) * \ldots * \mathbb{P}(A_n|A_1 \cap \ldots \cap A_{n-1})$$

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

Proof. Induct over the fact that $\mathbb{P}(A \cap B) = \mathbb{P}(A) * P(B|A)$. □

Example 1.3

Suppose we have 13 batteries, of which 3 are dead. Suppose we draw three batteries without replacement. What is the probability that all three are dead?

Let $D_i :=$ the event that the ith battery is dead, for $1 \leq i \leq 3$. Then we are interested in $\mathbb{P}(D_1 \cap D_2 \cap D_3)$. by the law of multiplication, we have

$$\mathbb{P}(D_1 \cap D_2 \cap D_3) = \mathbb{P}(D_1) * \mathbb{P}(D_2|D_1) * \mathbb{P}(D_3|D_1 \cap D_2) = \frac{3}{13}\frac{2}{12}\frac{1}{11} = \frac{1}{286}$$

In order to simplify problems, it is often beneficial to eliminate outcomes which cannot happen. For instance, consider the following example:

Example 1.4

You have 10 good batteries and 3 bad batteries. You pick out 4 good batteries and remove them. What is the probability that the fifth battery is good?

One approach is to notice that $\mathbb{P}(\geq 4 \text{ good batteries}) = 10/13 * 9/12 * 8/11 * 7/10$ and $\mathbb{P}(\geq 5 \text{ good batteries}) = 10/13 * 9/12 * 8/11 * 7/10 * 6/9$, and thus $\mathbb{P}(\geq 5| \geq 4) = \mathbb{P}(\geq 5 \cap \geq 4)/\mathbb{P}(\geq 4) = \mathbb{P}(\geq 5)/\mathbb{P}(\geq 4) = 6/9$. But a simpler way to do this to simply ignore the 4 good batteries that have been removed. If we reduce the sample space to the remaining 6 good batteries and 3 bad batteries, it is clear that the probability of picking a good battery is $6/9$.

Thus, we reduce the sample space after an event B has occurred by only considering those events where B occurs. We must also change our probability rule when we do this reduction:

Definition 1.7

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an event $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, the **reduction of sample space** by B is the probability space $(B, \mathcal{F} \cap B, \mathbb{P}_B)$, where $\mathcal{F} \cap B$ is the σ-algebra obtained by intersecting each element of $\mathcal{F}$ with B.

Next, we can consider the probabilities of the four possible configurations of A and B: that is, they both occur, neither occurs, only A occurs, or only B occurs. These probabilities are related by the law of total probability:

Theorem 1.3: Law of Total Probability

Let $A, B \in \mathcal{F}$, with $\mathbb{P}(B) > 0$. Then we have

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

More generally, suppose that $B_1, \dots B_n$ all have $\mathbb{P}(B_i) > 0$. Suppose also that the various B_i partition Ω (that is, $B_i \cap B_j = \varnothing$ for any $i \neq j$, and $\bigcup B_i = \Omega$). Then we have

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

Example 1.5

Suppose you randomly paint each side of a coin with heads or tails independently, with an equal probability of each. you flip the resulting coin and get heads. What is the probability you get heads again?

We are looking for $\mathbb{P}(HH|H)$. This is

$$\mathbb{P}(HH|H) = \frac{\mathbb{P}(HH \cap H)}{\mathbb{P}(H)}$$

But HH can only happen when H happens, so we can reduce the sample space to

$$\mathbb{P}(HH|H) = \frac{\mathbb{P}(HH)}{\mathbb{P}(H)}$$

Then we need to find $\mathbb{P}(H), \mathbb{P}(HH)$. Let 0, 1, 2 denote the events that the coin has the respective number of heads painted on it. Using the law of total probability:

$$\mathbb{P}(H) = \underbrace{\mathbb{P}(0)}_{1/4}\underbrace{\mathbb{P}(H|0)}_{0} + \underbrace{\mathbb{P}(1)}_{1/2}\underbrace{\mathbb{P}(H|1)}_{1/2} + \underbrace{\mathbb{P}(2)}_{1/4}\underbrace{\mathbb{P}(H|2)}_{1} = 1/2$$

Similarly,

$$\mathbb{P}(HH) = 1/4 * 0 + 1/2 * 1/4 + 1/4 * 1 = 3/8$$

So

$$\mathbb{P}(HH|H) = \frac{\mathbb{P}(HH)}{P(H)} = \frac{3/8}{1/2} = \frac{3}{4}$$

Lastly, the occurrence of one event may gives us new information about the state of a certain system, which changes the probabilities of the other events. In this case, we want to update our probabilities:

Theorem 1.4: Bayes' Theorem

Let $A, B \in \mathcal{F}$, with $\mathbb{P}(A) > 0$, $0 < \mathbb{P}(B) < 1$. Then

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

Proof. Apply the Law of Multiplication to the numerator and the Law of Total Probability to the denominator. □

In this case, B represents the probability of "before", and A is "after". In other words, suppose you initially believe there is a $\mathbb{P}(B)$ probability chance of B having happened (a priori). Afterwards, you observe that A has occurred. Then $\mathbb{P}(B|A)$ is our updated probability for B having happened (a posteriori), given that we know A has resulted.

Example 1.6

Suppose you have either a \$1 or \$20 bill in your pocket. You receive a \$1 bill in change and put it in your pocket. You then pull out a bill at random and find a \$1 bill. What is the probability that you had a \$1 bill originally?

Let A be the event that you pull out a \$1 bill, and B be the event that you originally had a \$1 bill. Using Bayes' Theorem,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} = \frac{1 * 1/2}{1 * 1/2 + 1/2 * 1/2} = \frac{1/2}{3/4} = \frac{2}{3}$$

Example 1.7

Suppose a test for a rare disease gives a false negative 5% of the time, and a false positive 2% of the time. Suppose only .1% of the population has this disease. Given that you receive a positive result, what is the probability you have the disease?

Let A be the event that the test returns positive, and B the event that you have the disease. Using Bayes' Theorem,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} = \frac{0.95 * 0.001}{0.95 * 0.001 + 0.02 * 0.999} \approx 0.04$$

So there is only a 4% chance of you actually having the disease, despite getting a positive test.

These examples show that intuitively, the reason that we need to update our credences in this way is that it is not necessarily equally likely that we start in B or B^c. To figure out

which set we started in, we need to find the relative probabilities that $B \cap A$ has occurred and that $B^c \cap A$ has occurred. This is what gives rise to the denominator in the theorem. One way to rewrite the formula to reflect this is

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

More generally, if we have disjoint $B_1, B_2, \ldots B_n$ which partition Ω, then we have a path through each of the B_i to get to A. So for any k, we would then have

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(B_k \cap A)}{\sum_i \mathbb{P}(B_i \cap A)} = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i)}$$

Example 1.8: The Monty Hall Problem

On a certain game show, three doors have prizes behind them. Two are goats and one is a million dollars. You go on the game show and choose a door. The host then opens one of the two remaining doors, revealing a goat. He offers you the option to switch to the other unopened door. Should you switch?

Suppose you adopt the strategy of switching. Since the host always shows you a goat, then only way you can lose by switching is if you initially pick the door with the million dollars. So you have a 2/3 chance of winning if you switch. This can be verified using Bayes' Theorem.

1.4 Independence

Suppose that we want to know the probability of an event A, and we receive the information that event B has happened. Then we should update our probability to $\mathbb{P}(A|B)$. But it may happen that B and A are completely unrelated, and that knowing B has occurred gives us no information about whether A has occurred. Then we say that these events are *independent*.

Definition 1.8

Two events $A, B \in \mathcal{F}$ with $\mathbb{P}(A), \mathbb{P}(B) > 0$ are **independent** if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

It is easily shown that A, B are independent if and only if $\mathbb{P}(B|A) = \mathbb{P}(B)$, so the definition is symmetric. Moreover, we can combine these two statements to show that they are independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A) * \mathbb{P}(B)$.

Lemma

If A, B are independent, then all of the following pairs are independent: (A, B^c), (A^c, B), (A^c, B^c).

Proof. For the first case, suppose A, B are independent. By the Law of Total Probability, $\mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) = \mathbb{P}(A)$. So $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$. By independence,

$\mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B^c)$. So A, B^c are independent. This then proves the other pairs, by relabeling A, B, A^c, B^c as necessary. $\square$

Note that A, B being disjoint does not mean that they are independent; in fact, disjoint sets are *always* independent (as long as $\mathbb{P}(A), \mathbb{P}(B) > 0$), since knowing one has occurred tells us that the other event has not occurred.

Example 1.9: The Jailer's Paradox

Alex, Ben, and Chris are in jail. The jailer learns that two of them will go free, and one will remain in jail, and he knows who. Alex asks the jailer to tell him the name of one of the other two men who will go free. The jailer refuses, and says that if he did so, then the probability that Alex would remain in jail would go from 1/3 to 1/2. Is the jailer correct?

let A, B, C be the events that Alex, Ben, and Chris get life imprisonment, respectively. Let J be the event that the jailer says Chris will go free (the situation is the same for Ben). Then by Bayes' Theorem, since A, B, C partition Ω,

$$\mathbb{P}(A|J) = \frac{\mathbb{P}(J|A) * \mathbb{P}(A)}{\mathbb{P}(J|A)\mathbb{P}(A) + \mathbb{P}(J|B)\mathbb{P}(B) + \mathbb{P}(J|C)\mathbb{P}(C)} = \frac{\frac{1}{2} * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + 1 * \frac{1}{3} + 0\frac{1}{3}} = \frac{1}{3}$$

So the probability does not change. Intuitively, this is because if Alex is staying imprisoned, then the jailer has free choice in who to tell Alex. If Alex is going free, then his choice is restricted.

Example 1.10

Suppose we flip 2 biased coins, which have a $p \in [0, 1]$ chance of landing heads. Assume each flip is independent.

Let H_i be the event that that ith flip is heads. By assumption, H_1, H_2 are independent, so $\mathbb{P}(H_1 \cap H_2) = \mathbb{P}(H_1)\mathbb{P}(H_2) = p \cdot p = p^2$. Since independence holds for complements, we have $\mathbb{P}(H_1 \cap H_2^c) = \mathbb{P}(H_1)\mathbb{P}(H_2^c) = p \cdot (1 - p)$, and similarly for the other combinations of events.

Note that we did not define the probability space explicitly. To do so, we would note that $\Omega = \{TT, HT, TH, HH\}$, and since we are able to distinguish all the outcomes, $\mathcal{F} = 2^\Omega$. Furthermore, it suffices to define the specific values of our probability rule:

$$\begin{cases} \mathbb{P}(\{TT\}) = (1 - p)^2 \\ \mathbb{P}(\{HH\}) = p^2 \\ \mathbb{P}(\{TH\}) = \mathbb{P}(\{HT\}) = p(1 - p) \end{cases}$$

Lastly, we can define the events $H_1 = \{HT, HH\}$ and $H_2 = \{TH, HH\}$. Then we can see that $\mathbb{P}(H_1) = \mathbb{P}(\{HH\}) + \mathbb{P}(\{HT\}) = p^2 + p(1 - p) = p$, and the same for H_2. Then we can verify that $\mathbb{P}(H_1 \cap H_2) = \mathbb{P}(\{HH\}) = p^2 = p \cdot p = \mathbb{P}(H_1)\mathbb{P}(H_2)$. So owe see that we indeed have independent events.

Example 1.11: Paradox of Pairwise Independence (Bernstein)

Suppose you have three events $A, B, C \in \mathcal{F}$. Suppose that each pair of events is independent. Then are all three independent? That is, knowing the outcome of one event doesn't tell you anything about the other two. Does knowing the outcome of two events tell you anything about the last one?

Using the biased coins from the previous example, let $A = H_1, B = H_2, C = \{HT, TH\}$. Suppose we let $p = 0.5$. We already know that A, B are independent, and it is easily verified that A, C and B, C are independent when $p = 0.5$. But suppose we know that A and B have both occurred. Then we must have $H_1 \cap H_2 = \{HH\}$. But this is disjoint with C, so C cannot happen. Thus we conclude that pairwise independence is not sufficient to conclude independence of all the events.

The last example suggested the notion of multiple events being independent simultaneously.

Definition 1.9

We say that events $A_1, A_2, \ldots, A_n$ are **independent** if for any i and any $J \subseteq \{1, 2, \ldots, n\}$ with $i \notin J$, we have $\mathbb{P}(A_i | \bigcap_{j \in J} A_j) = \mathbb{P}(A_i)$. In other words, the occurrence of any combination of the other events tells us nothing about the probability of A_i.

Similarly, to the definition of independence for two events, we can find a more symmetric definition of independence for multiple events:

Theorem 1.5

A finite collection of events $A_1, A_2, \ldots, A_n$ are independent if and only if, for any $J \subseteq A$, $\mathbb{P}(\bigcap_{j \in J} A_j) = \prod_{j \in J} \mathbb{P}(A_j)$.

Proof. ($\Longrightarrow$) Label our A_j as $A_{j_1}, A_{j_2}, \ldots, A_{j_k}$, $1 \leq k \leq n$. Then by the law of multiplication,

$$\mathbb{P}(A_{j_1} \cap A_{j_2} \cap \ldots \cap A_{j_k}) = \mathbb{P}(A_{j_1})\mathbb{P}(A_{j_2}|A_{j_1}) \ldots \mathbb{P}(A_{j_k}|A_{j_1} \cap A_{j_2} \cap \ldots A_{j_{k-1}})$$

But we assume independence, so this reduces to

$$\mathbb{P}(A_{j_1} \cap A_{j_2} \cap \ldots \cap A_{j_k}) = \mathbb{P}(A_{j_1})\mathbb{P}(A_{j_2}) \ldots \mathbb{P}(A_{j_k})$$

($\Longleftarrow$) We skip this proof. □

Example 1.12

Suppose we flip 5 biased p-coins, and the tosses are independent. What is the probability of 4 heads and 1 tails?

Letting H_i be the event that the ith flip is heads. Then we observe that

$$\begin{gathered}\mathbb{P}(4 \text{ heads, } 1 \text{ tails}) = \\ \mathbb{P}(H_1^c H_2 \dots H_5) + \mathbb{P}(H_1 H_2^c H_3 \dots H_5) + \dots + \mathbb{P}(H_1 \dots H_4 H_5^c) \\ = 5p^4(1-p)\end{gathered}$$

Chapter 2

Random Variables

2.1 Random Variables

In order to quantify the outcomes of experiments, it is helpful to be able to define *random variables* that represent the result of an experiment with more precision than the binary paradigm of an event happening or not happening. We can do this by considering functions of the form $X : \Omega \to \mathbb{R}, \omega \mapsto X(\omega)$.

Definition 2.1

Given $A \subseteq \Omega$, the **indicator function** of A is

$$1_A(\omega) \coloneqq \begin{cases} 1, \omega \in A \\ 0, \omega \notin A \end{cases}$$

In other words, given an outcome ω, $1_A(\omega)$ indicates whether or not A has occurred.

Example 2.1

Suppose we roll a die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$, and we can define $X(\omega) = \omega$.

Example 2.2

Suppose we roll two dice. Then $\Omega = \{(i, j) | 1 \leq i, j \leq 6\}$. Suppose we want to know the sum of the dice. Then we define $X(\omega) = i + j, \omega = (i, j)$.

Definition 2.2

Given a function $X : \Omega \to \mathbb{R}$, we define the **range** of values $R_X = \{X(\omega) : \omega \in \Omega\}$. We define the set $\{X = a\}$ for $a \in \mathbb{R}$ as $\{\omega \in \Omega : X(\omega) = a\}$. This allows us to write, for instance, $\mathbb{P}(X = a)$. We define $\{a < X < b\}$ and other sets of that form similarly.

Definition 2.3

Given a random variable $X : \Omega \to \mathbb{R}$, $\sigma(X)$ is the smallest $\sigma - algebra$ containing all sets of the form $\{a < X < b\}$ for any $a, b \in \mathbb{R}$ (or the σ-algebra generated by all sets of that form). In other words, $\sigma(X)$ quantifies the information obtained when we know X takes on a certain value.

Definition 2.4

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, an **$\mathcal{F}$-random (real) variable** is a function $X : \Omega \to \mathbb{R}$ such that $\sigma(X) \subseteq \mathcal{F}$. Intuitively, this means you can observe the random variable X by running an experiment in $\mathcal{F}$.

Example 2.3

Suppose $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $\mathbb{P}(\omega_i) = 1/6$. Then suppose we define $X : \Omega \to \mathbb{R}$ to be $1_{\{1,2\}}(\omega)$. Then X is an $\mathcal{F}$-random variable on $\mathcal{F} = 2^\Omega$. But suppose we define $\mathcal{G} = \{\{1, 2, 3\}, \{4, 5, 6\}, \varnothing, \Omega\}$. Then X is not a $\mathcal{G}$-random variable, since we are not able to separate sets based on $1_{\{1,2,\}}(\omega)$.

In other words, we must be able to conclusively know the value of X if we get the right outcome in $\mathcal{F}$ (or $\mathcal{F}$ separates values of X). Note that if $\mathcal{F} = 2^\Omega$, then any function $X : \Omega \to \mathbb{R}$ is a random variable.

At the moment, we have placed very few restrictions on the actual values that X may assume. In particular, it may take values all along the real line, in a finite set (such as the indicator function), or a countably infinite set. We will first consider sets which take values in a countable set (finite or countably infinite)¿

Definition 2.5

An $\mathcal{F}$-random variable $X : \Omega \to \mathbb{R}$ is **discrete** if R_X is countable. In this case, then $\sigma(X)$ is generated by the countable sets of the form $\{X = x\}, x \in R_x$, and every set of the form $\{a < X < b\} = \bigcup_{a<x<b, x \in R_x} \{X = x\}$.

Definition 2.6

Given a discrete random variable $X : \Omega \to \mathbb{R}$, the **probability mass function** of X is $f_X(x) : R_x \to [0, 1]$, with $f_X(x) \coloneqq \mathbb{P}(X = x)$

Example 2.4

Suppose we roll a die and define $X(\omega) = \alpha 1_{\{1,2\}}(\omega + \beta 1_{\{3,4\}}(\omega)$ (with $\alpha \neq \beta \neq 0$, so that we can distinguish between $\{1,2\}, \{3,4\}, \{5,6\}$. Then our σ-algebra is simply the algebra generated by the sets we can distinguish between; namely, $\sigma(X) = \sigma(\{\{1,2\}, \{3,4\}, \{5,6\}\}$. Here, X is a discrete variable.

Alternatively, we can also consider functions that take values in an interval:

Definition 2.7

An $\mathcal{F}$-random variable $X : \Omega \to \mathbb{R}$ is **continuous** if there exists a function $f_X : \mathbb{R} \to [0, \infty)$ such that $\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$. In this case, f_X is called the **probability density function** of X.

Example 2.5

Suppose X is chosen "uniformly at random" in $[\alpha, \beta]$, with $\alpha < \beta$. Then we must have $f_X(x) = \frac{1}{\beta - \alpha} 1_{\{\alpha,\beta\}}(x)$.

Note that the above example demonstrates that X need not actually take values in all of $\mathbb{R}$ in order to define $f_X : \mathbb{R} \to [0, \infty)$. Moreover, note that probability mass functions are defined for discrete variables, while probability density functions are defined for continuous variables. In both cases, these functions represent the "instantaneous probability" of a certain value of X.

Definition 2.8

Given an $\mathcal{F}$-random variable X, the **distribution** of X is the probability rule $\mathbb{P}_X$, defined on $(\mathbb{R}, \beta_{\mathbb{R}})$ by $\mathbb{P}_x(E) \coloneqq \mathbb{P}(X \subseteq E)$, $E \in \beta_{\mathbb{R}}$. Recall that $\beta_{\mathbb{R}}$ is the σ-algebra generated by open intervals of $\mathbb{R}$.

We should note that two different random variables may have the same distribution. For instance, the number of heads and number of tails are different variables with identical distributions.

Example 2.6

Suppose we roll a die and consider the indicator $X = 1_{\{1,2\}}$. Then the distribution $\mathbb{P}_X$ is defined as $\mathbb{P}_X(X = 0) = 2/3$, $\mathbb{P}_X(X = 1) = 1/3$.

2.2 Expected Value

In many cases, we would like to know what we can expect the value of a variable X is, before we have actually observed any experiment. In a sense, we want to know the average of X. If

we were to consider this value by taking the limit of outcomes of an experiment, multiplying each value by the proportion of experiments in which it occurs and then summing the values. This suggests that, at least for a discrete variable, the expectation should be

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

Using the language of pmfs and pdfs, we can write this in a concise way that reflects correspondence between discrete and continuous random variables:

Definition 2.9

Given an $\mathcal{F}$-random variable X, the **expectation** of X is

$$\mathbb{E}[X] := \begin{cases} \sum_{x \in R_X} x f_X(x), X \text{ is discrete} \\ \int_{R_X} x f_X(x) dx, X \text{ is continuous} \end{cases}$$

One particularly important fact is that for any event A, the expected value of the indicator function 1_A is simply

$$\mathbb{E}[1_A] = 1 * \mathbb{P}(A) + 0 * \mathbb{P}(A^c) = \mathbb{P}(A)$$

Example 2.7

Suppose we want to let the random variable X be the the number of flips required before getting heads (formally, Ω is the set of all infinite sequences of H, T). Then X is discrete, so we define the pmf $f_X(n) = \mathbb{P}(X = n)$ as the probability of $n - 1$ tails, and then a heads. Assuming the flips are independent with a probability p of heads, then $f_X(n) = (1 - p)^{n-1}p$.

Definition 2.10

We say that an $\mathcal{F}$-random variable X has a **geometric p-distribution** if its mass function is given by $(1 - p)^{x-1}p$, for $x \geq 1$.

More generally, we can consider "random elements," which are essentially random variables that are not required to take real values; in other words, functions of the form $Z : \Omega \to \mathcal{D}$.

Example 2.8

Suppose we pick a random card from a deck. Then its suit is a random element, $Z : \Omega \to \mathcal{D}$, where $\mathcal{D} = \{H, C, S, D\}$.

Similarly, we could think of random elements into the set $\mathbb{R}^2$, where each coordinate is a random variable $X, Y : \Omega \to \mathbb{R}$. However, one issue with using random elements is that we can no longer take an expectation. To resolve this, we can add a quantifying function

$g : \mathcal{D} \to \mathbb{R}$. Then the function $g \circ Z : \Omega \to \mathbb{R}$ is a random variable. This allows us to generalize results about random variables to many kinds random elements.

Theorem 2.1: Law of the Unconscious Statistician

Given a random element Z and a function $g : \mathcal{Z} \to \mathbb{R}$, we have

$$\mathbb{E}[g(z)] = \lim_{n\to\infty} \frac{1}{N} \sum_{k=1}^{N} g(z_k)$$

Then if g is discrete, we can say:

$$\lim_{n\to\infty} \frac{1}{N} \sum_{k=1}^{N} g(z_k) = \lim_{n\to\infty} \frac{1}{N} \sum_{k=1}^{N} \sum_{z\in R_Z} g(z) 1_{\{z_k\}}(z)$$

$$= \sum_{z\in R_Z} g(z) \lim_{n\to\infty} \left(\frac{\sum_{k=1}^{N} 1_{z_k}(z)}{N} \right) = \sum_{z\in R_z} g(z)\mathbb{P}(Z = z)$$

If g is continuous, we simply replace $\sum_{z\in R_Z}$ with $\int_{R_Z}$:

$$\mathbb{E}[g(Z)] = \begin{cases} \sum_{z\in R_Z} g(z) f_Z(z), & \text{Z discrete} \\ \int_{R_Z} g(z) f_Z(z), & \text{Z continuous} \end{cases}$$

In other words, the expectation of a random variable can be computed as the limit of independent trials.

Of particular importance is the class of random variables that combines multiple random variables into an n-tuple.

Definition 2.11

Given a random element $Z : \Omega \to \mathbb{R}^2$, with $Z \coloneqq (X, Y)$ for two random variables $X, Y : \Omega \to \mathbb{R}$, the **joint mass function** of X, Y is $f_Z(z) = f_{(X,Y)}(x, y) \coloneqq \mathbb{P}(X = x, Y = y)$. Similarly, the **joint distribution** is $\mathbb{P}_{(X,Y)}(E) \coloneqq \mathbb{P}(X, Y) \in E$. This definition can be extended to a set of random variables $X_1, X_2, \ldots, X_n$.

Note that it is *not* enough to know each of the individual probability distributions $\mathbb{P}_X, \mathbb{P}_Y$ in order to find the joint distribution. However, we can still find certain information about the joint distribution given information about the individual distributions.

Theorem 2.2: Linearity of Expected Value

Given two random variables X, Y, we have

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Proof. Let $Z = (X, Y)$, $g(z) = g(x, y) = x + y$. Then

$$\mathbb{E}[X+Y] = \mathbb{E}[g(z)] = \sum_{z \in R_Z} g(z)\mathbb{P}(Z - z) = \sum_{x \in R_X, y \in R_Y} (x+y)\mathbb{P}(X = x, Y = y)$$

$$= \sum_{x \in R_X} x \left(\sum_{y \in R_Y} \mathbb{P}(X = x, Y = y) \right) + \sum_{y \in R_Y} y \left(\sum_{x \in R_X} \mathbb{P}(X = x, Y = y) \right)$$

$$= \sum_{x \in R_x} x\mathbb{P}(X = x) + \sum_{y \in R_Y} y\mathbb{P}(Y = y) = \mathbb{E}[X] + \mathbb{E}[Y]$$ □

Example 2.9

At a party with n people, how many pairs of people to you expect to share a birthday?

Consider each person separately. Let X_i be the birthday of the ith person. Define $Z = \sum_{1 \le i < j \le n} 1_{X_i = X_j}$. Let $Y_{ij} = 1_{X_i = X_j}$. Then by the linearity of expected value,

$$\mathbb{E}[Z] = \sum_{i<j} \mathbb{E}[Y_{ij}]$$

Note that for any random variable,

$$\mathbb{E}[1_Y] = \sum_{y \in R_Y} y\mathbb{P}(Y = y) = 0 * \mathbb{P}(1_Y = 0) + \mathbb{P}(1_Y = 1) = \mathbb{P}(Y)$$

Now for any given $i < j$, $\mathbb{P}(X_i = X_j) = 1/365$. So

$$\mathbb{E}[Z] = \sum_{i<j} 1/365 = \frac{n}{2}\frac{1}{365} = \frac{n(n+1)}{2}\frac{1}{365}$$

The above example illustrates that two variables which are not independent still have the property of linearity. Thus, linearity of expected value is an extremely powerful result that applies to all random variables.

Definition 2.12

We say two random variables X, Y are **independent** if, for all $x \in R_X, y \in R_Y$, we have $f_{(X,Y)}(x, y) = f_X(x)f_Y(y)$.

Theorem 2.3

Two discrete random variables X, Y are independent if and only if each pair $\{X = x\}, \{Y = y\}$ for $x \in R_X, y \in R_Y$ are independent events.

Proof. If this is the case, then $f_{(X,Y)}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) = f_X(x), f_Y(y)$ for any $x \in R_X, y \in R_Y$. □

Theorem 2.4

Given two random elements X, Y and quantifiers g, h, if X, Y are independent, then $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$.

2.3 Conditional Expectation

Just as the occurrence of an event may update our probabilities of a certain event occurring, the occurrence of an event may also update our understanding of the distribution of a random variable.

Recall that given that an event B with $\mathbb{P}(B) > 0$ has occurred, we can reduce the sample space to $(B, \mathcal{F} \cap B, \mathbb{P}_B)$, where $\mathbb{P}_B(A) = \mathbb{P}(A|B)$. We might similarly define the conditional expectation of a random variable X. If we know an event B with $\mathbb{P}(B) > 0$ has occurred, we might define the conditional expectation of a discrete variables as

$$\mathbb{E}[X|B] = \sum_{x \in R_X} x\mathbb{P}_B(X = x) = \sum_{x \in R_X} x\mathbb{P}(X = x|B)$$

Example 2.10

Suppose we roll a die and are told that $B = \{1, 2, 3\}$ has occurred. Then $\mathbb{E}_B[X] = 2$ and $\mathbb{E}_{B^c}[X] = 5$.

However, we can rewrite this in an alternative form. Note that if $x \neq 0$ and $\{X = x\} \cap B$ is nonempty, then we have

$$\{X = x\} \cap B = \{X * 1_B = x\}$$

On the other hand, if $x = 0$ or if $\{X = x\} \cap B$ is empty, then

$$x\mathbb{P}(\{X = x\} \cap B) = 0$$

Thus we can write

$$\mathbb{E}[X|B] = \sum_{x \in R_X} x\frac{\mathbb{P}(\{X = x\} \cap B}{\mathbb{P}(B)} = \sum_{x \in R_X \cap B} x\frac{\mathbb{P}(X1_B = x)}{\mathbb{P}(B)} = \frac{\mathbb{E}[X1_B]}{\mathbb{P}(B)}$$

where the last equality is given by the law of the unconscious statistician. Then we can write an alternate equation for $\mathbb{E}[X|B]$ in order to generalize to all kinds of random variables.

Definition 2.13

Given any random variable X and any event B with $\mathbb{P}(B) > 0$, define the **conditional expectation** of X to be

$$\mathbb{E}[X|B] = \frac{\mathbb{E}[X1_B]}{\mathbb{P}(B)}$$

In the discrete case, this corresponds with the elementary formula we found above:

$$\mathbb{E}[X|B] = \sum_{x \in R_X} x\mathbb{P}(X = x|B)$$

Then suppose we want to know how the value of another random variable Y updates our expectation of X, without knowing the exact value of Y. In other words, we consider events of the form $B = \{Y = y\}$. Then we have

$$\mathbb{E}[X|Y = y] = \frac{\mathbb{E}[X1_{(Y=y)}]}{\mathbb{P}(Y = y)} = \sum_{x \in R_X} x\frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

Notice that $\mathbb{P}(X = x, Y = y)$ is simply $f_{(X,Y)}(x, y)$, and $\mathbb{P}(Y = y)$ is $f_Y(y)$. So we can write

$$\mathbb{E}[X|Y = y] = x\frac{f_{(X,Y)}(x, y)}{f_Y(y)}$$

This allows to generalize nicely to the continuous case:

Definition 2.14

Given two random variables X, Y, define the **conditional joint density function** as

$$f_{X|Y}(x|y) \coloneqq \frac{f_{(X,Y)}(x, y)}{f_Y(y)}$$

Example 2.11

Suppose we let X be the outcome of a die roll, and let $Y = 1_{\{1,2,3\}}$. Then we can write the expectation of X given Y given some event ω as

$$\mathbb{E}[X|Y](\omega) = 2 * 1_{\{1,2,3\}}(\omega) + 5 * 1_{\{4,5,6\}} = 2Y + 5(1 - Y)$$

Note that in the above example, we could scaled Y by π, e, or any other nonzero scalar without changing the situation. Thus, we arrive at the important observation that the actual value of Y doesn't matter; only the set of what events it allows us to conclude has occurred.

Suppose we have discrete random variables X, Y. Then the conditional expectation of X given Y is a function $\mathbb{E}[X|Y] : \Omega \to \mathbb{R}$ defined by

$$\mathbb{E}[X|Y](\omega) \coloneqq \mathbb{E}[X|Y = Y(\omega)]$$

Alternatively, define $\psi(y) \coloneqq \mathbb{E}[X|Y=y]$. Then $\mathbb{E}[X|Y](\omega) = \psi(Y(\omega) = \mathbb{E}[X|Y=Y(\omega)]$. Note that $\mathbb{E}[X|Y=Y(\omega)] = \sum_{x\in R_X} x f_{X|Y}(x|Y(\omega))$. Once again, this gives a nice method to generalize to the continuous case.

Definition 2.15

Given two random variables X, Y, the **conditional expectation** of X given Y is a function $\mathbb{E}[X|Y] : \Omega \to \mathbb{R}$ given by

$$\mathbb{E}[X|Y](\omega) \coloneqq \begin{cases} \sum_{x\in R_X} x f_{X|Y}(x|Y(\omega)) \\ \int_{R_X} x f_{X|Y}(x|Y(\omega)) \end{cases}$$

Remark

Note that $\mathbb{E}[X|Y]$ is no longer a single value, but a function that gives a value for each individual value of Y. This makes sense, since without knowing ahead of time what Y is, we would otherwise just get $\mathbb{E}[X]$.

As we will see, many of the theorems from conditional probability of random events hold for conditional expectation for random variables.

Theorem 2.5: Law of Total Expectation

Given any variables X, Y, we have

$$\mathbb{E}[X] = \mathbb{E}_1[\mathbb{E}_2[X|Y]] = \sum_{y\in R_Y} \mathbb{E}[X|Y=y]\mathbb{P}(Y=y)$$

Note here that $\mathbb{E}_2$ is a function that gives the "subaverage" of X given a certain value of Y. In this case, the interpretation here is that $\mathbb{E}_1$, the average value of X, is the weighted average of each of the subgroups given values of Y.

Example 2.12

Suppose we draw digits 0-9 independently. Suppose we let X be the number of draws until we get three 0s in a row. What is $\mathbb{E}[X]$?

Suppose we let Y be the number of draws until the first nonzero digit. Then by the law of total probability, we have

$$\mathbb{E}[X] = \sum_{y=1}^{\infty} \mathbb{E}[X|Y-y]\mathbb{P}(Y=y)$$

Suppose $y = 1, 2, 3$. Then we essentially just "restart" and increase the expectation

by y. If $y \geq 4$, then X has happened in 3 steps So we have

$$\mathbb{E}[X] = \sum_{y=1}^{3}(y + \mathbb{E}[X])\mathbb{P}(Y = y) + \sum_{y=4}^{\infty} 3\mathbb{P}(Y = y)$$
$$= (1 + \mathbb{E}[X])\frac{9}{10} + (2 + \mathbb{E}[X])\frac{9}{100} + (3 + \mathbb{E}[X])\frac{9}{1000} + 3\sum_{y=4}^{\infty}\frac{9}{10^y}$$
$$= \frac{999}{1000}\mathbb{E}[X] + \frac{1107}{1000} + 0.001$$
$$\implies \mathbb{E}[X] = 1110$$

In the above, we saw that we can treat the expectation of a random variable X conditional on a random variable Y as simply being a function mapping different events to the conditional expectation of X given a certain value of Y. Then using this interpretation, we can similarly define the conditional probability of an event given a variable, using the observation earlier that $\mathbb{P}(A) = \mathbb{E}[1_A]$ for any event A.

Definition 2.16

Given an event A and a random variable Y, the conditional probability of A given Y is a function $\mathbb{P}(A|Y) : \Omega \to [0, 1]$ defined by

$$\mathbb{P}(A|Y)(\omega) := \mathbb{E}[1_A|Y](\omega)$$

Example 2.13

Let Z be the sum of outcomes of two dice. Let X, Y, be the first and second rolls, respectively, and B the event that the first roll is 1. Then $\mathbb{P}(Z = k|B) = 0$ if $k > 7$, and

$$\mathbb{E}[Z|B] = \sum_{k=2}^{7}\mathbb{P}(Z = k|B) = \frac{1}{6}\sum_{k=2}^{7} k = \frac{9}{2} = 4.5$$

Alternatively, note that $Z = X+Y$. Since we know $X = 1$, and X, Y are independent, we have

$$\mathbb{E}[Z|X] = \mathbb{E}[X|X] + \mathbb{E}[Y|X] = X + \mathbb{E}[Y]$$

Here are some basic properties of conditioning with random variables:

1. If X, Y are independent, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.
2. If X is completely dependent on Y (such that $X = f(Y)$), then $\mathbb{E}[X|Y] = f(Y) = X$.
3. Regardless of dependence between X, Y, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

One way to interpret these properties is to note that they are all of the form $\mathbb{E}[g(X, Y)|Y]$. For instance, in the first case, we have $g(X, Y) = X$, in the second we have $g(X, Y) = X =$

$f(Y)$, and in the third we have $g(X,Y) = X + Y$. Then we can expand:

$$\mathbb{E}[g(X,Y)|Y) = \sum_{x \in R_X} g(x,Y) f_{X|Y}(x|Y)$$

This idea leads to the following additional properties:

4 If $g(X,Y) = f(X)h(Y)$, then $\mathbb{E}[f(X)h(Y)|Y] = h(Y)\mathbb{E}[f(X)|Y]$.

5 $\mathbb{E}[g(X,Y)|Y](y) = \mathbb{E}[g(X,y)]$

Example 2.14: Buffon's Needle

Suppose you drop a needle of length l onto an infinite paper with ruled vertical lines at intervals of d, with $l < d$. What is the probability the needle crosses a line?

Let X be the distance of the center to the nearest vertical line. Then $X \sim \text{Uniform}(0, d/2)$. Let θ be the angle of the line determined by the needle with the nearest ruled line. Then $\theta \sim \text{Uniform}(0, \pi/2)$. Suppose we draw a triangle from the intersection of the vertical line and needle line, and the altitude dropped from the midpoint of the needle to the vertical line. Then the hypotenuse is $x/\sin\theta$, and the needle crosses the vertical line if and only if $x/\sin\theta < l/2$. Then

$$\mathbb{P}(H < l/2) = \mathbb{P}(X < l\sin\theta/2) = \mathbb{E}[\mathbb{P}(X < l\sin\theta/2|\theta)]$$

Since X, θ are independent, we have

$$\mathbb{E}[\mathbb{P}(X < l\sin\theta/2)] = \mathbb{E}\left[\frac{l\sin\theta}{d}\right] = \frac{l}{d}\int_0^{\pi/2} \sin\theta \frac{1}{\pi/2} d\theta = \frac{2}{\pi}\frac{l}{d}$$

Summary

Types of Conditional Expectation

Given $\mathcal{F}$-random variables X, Y, and a random event A, we have:

- $\mathbb{E}[X|A] = \frac{\mathbb{E}[X 1_A]}{\mathbb{P}(A)}$
- $\mathbb{E}[X|Y] : \omega \to \mathbb{R}$ is an $\mathcal{F}$-random variable with $\mathbb{E}[X|Y](\omega) \coloneqq \mathbb{E}[X|Y = Y(\omega)]$.
- $\mathbb{P}(A|X) : \omega \to \mathbb{R}$ is an $\mathcal{F}$-random variable defined by

$$\mathbb{P}(A|X)(\omega) \coloneqq \mathbb{E}[1_A|X](\omega) = \mathbb{E}[1_A|X = X(\omega)] = \mathbb{P}(A|X = X(\omega))$$

Basic Properties of Conditional Expectation

1. If X, Y are independent, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.
2. If $X = f(Y)$, then $\mathbb{E}[X|Y] = f(Y) = X$.
3. Regardless of dependence between X, Y, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
4. If $g(X, Y) = f(X)h(Y)$, and Y is a $\mathcal{G}$-random variable with $X \perp \mathcal{G}$,

$$\mathbb{E}[f(X)h(Y)|Y] = h(Y)\mathbb{E}[f(X)|Y]$$

5. $\mathbb{E}[g(X, Y)|Y](y) = \mathbb{E}[g(X, y)]$.
6. $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \sum_{R_X} \mathbb{E}[X|Y] f_Y(y)$.

Chapter 3

Stochastic Processes

3.1 Bernoulli Processes

Suppose we flip a biased p-coin. Then if we let $X = 1$ when heads is flipped and $X = 0$ when tails is flipped, we say that X follows a *Bernoulli distribution* with parameter p; that is, it takes the value of 1 with probability p and the value of 0 with probability $q = 1 - p$.

Definition 3.1

We say that a random variable X follows the **Bernoulli distribution** with parameter p if if only takes the values $0, 1$, with $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. In this case, we write $X \sim \text{Bernoulli}(p)$.

Now suppose we repeatedly flip this biased coin and consider the sequence of outcomes $X_1, X_2, \ldots$. Then these variables are independent and identically distributed. Moreover, we can interpret this collection of trials as a process, roughly meaning that there is a sense of progression or time. Thus, a set of Bernoulli trials is an example of what we call a stochastic (or random) process.

For the time being, we will informally define a stochastic process as a collection of random variables $\{X_t\}_{t\in\tau}$, where t essentially represents time, and τ is an indexing set that is either $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ or $[0, \infty)$. Then the collection $(X_n)_{n\geq 0}$ is a stochastic processes that we call a Bernoulli process.

Definition 3.2

A **Bernoulli process** with parameter p and N trials is a collection of random variables X_i, $1 \leq i \leq N$, such that each $X_i \sim \text{Bernoulli}(p)$ and the trials are independent. (In other words, they are **independent and identically distributed** (iid))

Suppose we define S_n as the number of successes in the first n flips:

$$S_n \coloneqq \sum_{k=1}^{n} X_k (n \geq 1),\ S_0 = 0,\ S_{k+1} = X_{k+1} + S_k$$

Then by the binomial formula, we have

$$f_{S_n}(k) = \mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k},\ k \in \mathbb{N} \implies S_n \sim \text{Bin}(n, p)$$

Moreover, by the linearity of expected value, we have

$$\mathbb{E}[S_n] = \sum_{k=1}^{n} \mathbb{E}[X_k] = np$$

Now suppose we are given $n, m \in \mathbb{N}$. Then what is the distribution of the number of flips in the time period $(n, n+m]$? In other words, what is the distribution of $S_{n+m} - S_n$? In general, we would need the joint distribution to determine this, but because a Bernoulli process consists of independent variables, we can easily calculate this with

$$S_{n+m} - S_n = X_{n+1} + X_{n+2} + \ldots + X_{n+m} = \sum_{k=1}^{m} X_{n+k} \sim \text{Bin}(m, p)$$

In particular, note that this distribution is *stationary*, meaning that $S_{n+m} - S_n \sim S_m$ for any $n \geq 0$. It also has the property of *independent increments*, meaning that $S_m - S_n$ is independent of $S_k - S_l$ for any $m \geq n, k \geq l$ (as long as the intervals don't overlap).

Then if we construct the joint distribution $f_{(S_n, S_{n+m})}(k, l)$, we get

$$\begin{aligned}
f_{(S_n, S_{m+n})}(k, l) &\coloneqq \mathbb{P}(S_n = k, S_{n+m} = l) \\
&= \mathbb{P}(S_n = k, S_{n+m} - S_n = l - k) \\
\text{(Independent increments)} &= \mathbb{P}(S_n = k)\mathbb{P}(S_{n+m} - S_n = l - k) \\
\text{(Stationary)} &= \mathbb{P}(S_n = k)\mathbb{P}(S_m = l - k) \\
&= \binom{n}{k} p^k (1-p)^{n-k} * \binom{m}{l-k} p^{l-k} (1-p)^{m-(l-k)}
\end{aligned}$$

So we find that the joint distribution is given by

$$f_{(S_n, S_{n+m})}(k, l) = \binom{n}{k}\binom{m}{l-k} p^l (1-p)^{n+m-l}, 0 \leq k \leq l \leq n+m, l-k \leq m$$

We next consider the "arrival time" of a Bernoulli process. Define T_k to be the time at which a Bernoulli process reaches k successes. Formally,

$$T_k \coloneqq \min\{n : S_n = k\}$$

Then the distribution is given by

$$f_{T_k}(n) = \mathbb{P}(S_{n-1} = k-1, X_n = 1) = \underbrace{\mathbb{P}(S_{n-1} = k-1)\mathbb{P}(X_n = 1)}_{\text{Independent increments}}$$

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

In particular, the variable T_1 has distribution $f_{T_1}(k) = (1-p)^{k-1}p$, which we denote Geom($p$). Moreover, we can see that

$$1 = \sum_{k=1}^{\infty} f_{T_1}(k) = \sum_{k=1}^{\infty}(1-p)^{k-1}p = \sum_{k=0}^{\infty} q^k(1-q)$$

Rearranging, we get the geometric series formula

$$\frac{1}{1-q} = \sum_{k=0}^{\infty} q^n,\ q \in (-1,1)$$

This allows us to calculate the expectation of T_1, or the number of trials expected before flipping a heads.

$$\mathbb{E}[T_1] = \mathbb{E}[\mathbb{E}[T_1|X_1]] = p\mathbb{E}[T_1|X_1=1] + (1-p)\mathbb{E}[T_1|X_1=0] = p + (1-p)(1+\mathbb{E}[T_1])$$
$$\implies p\mathbb{E}[T_1] = 1 \implies \mathbb{E}[T_1] = \frac{1}{p}$$

Similarly to asking the time before flipping the first heads, we can also ask more generally how long it will take between the kth and $k+1$th heads, or the "interarrival time." If we let T_k represent the number of flips until the kth heads, then we can calculate

$$\mathbb{P}(T_1 = n_1, T_2 - T_1 = n_2, \dots T_k - T_{k-1} = n_k) = q^{n_1-1}p * q^{n_2-1}p \dots q^{n_k-1}p = \prod_{i=1}^{k} \mathbb{P}(T_1 = n_i)$$

Thus we see that the interarrival times are independent and identically distributed.

Theorem 3.1

Let $(X_i)_{i\geq 1}$ be a Bernoulli process, and define $T_k = \min\{n : S_n = \sum_i^n X_i = k\}$. Then the collection $(T_k)_{k\geq 1}$ has the following properties:

- (Independent) Each $T_k - T_{k-1}$ is independent from the others.
- (Identically distributed) Each $T_k - T_{k-1} \sim T_1 \sim$ Geom(p).

A basic calculation also allows us to find the expected value of T_k:

$$\mathbb{E}[T_k] = \mathbb{E}[(T_k - T_{k-1}) + (T_{k-1} - T_{k-2}) + \dots + (T_2 - T_1) + T_1] = \underbrace{1/p + 1/p + 1/p \dots + 1/p}_{k \text{ times}} = \frac{k}{p}$$

3.2 Poisson Processes

The Bernoulli process allows us to model random occurrences over discrete time. In order to model a situation of random arrivals over continuous time, we can use the Poisson distribution, which is essentially the infinite limit of the binomial distribution.

Suppose we let N_1 be the number of arrivals in 1 hour. Suppose that these arrivals are random and independent, with equal probability at every time, and have an average occurrence of $\lambda > 0$ per hour, so that $\mathbb{E}[N_1] = \lambda$. Then what is the distribution of N_1?

Since we know that the probability of an arrival is the same "at every time," we can approximate the continuous distribution by dividing the hour into n equal subintervals. We can then ask whether each subinterval has an arrival or not. If we let X_i be the variable representing the ith subinterval, then $X_i \sim \text{Bernoulli}(p)$, and we approximate N_1 with $S_n = \sum_{k=1}^{n} X_k \sim \text{Bin}(n, p)$. Moreover, we have $\lambda = \mathbb{E}[N_1] \approx \mathbb{E}[S_n] = np_n$. Then we have $p_n = \lambda/n$ (which is guaranteed to be in $[0, 1]$ for large n).

Plugging back into the approximation, we have

$$\begin{aligned} f_{N_1}(k) = \mathbb{P}(N_1 = k) &\approx \mathbb{P}(S_n = k) \\ &= \binom{n}{k} p^k (1-p)^{n-k} \\ = \frac{\lambda^k}{k!} \underbrace{\frac{n}{n} \frac{(n-1)}{n} \cdots \frac{n-(k+1)}{n}}_{\to 1} &\underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\to 1} \end{aligned}$$

Then if we take the limit as our approximation becomes finer, we get

$$f_{N_1}(k) = \mathbb{P}(N_1 = k) = \lim_{n \to \infty} \mathbb{P}(S_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Thus we have derived the Poisson distribution.

Definition 3.3

A random variable X is **Poisson** with parameter λ, or $X \sim \text{Poisson}(\lambda)$, if

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

For the general case, where the average rate of arrivals is λ per hour, and we want to know the number of arrivals in t hours, we can simply plug this into the distribution, and find that $N_t \sim \text{Poisson}(\lambda t)$, such that

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

Since we derived the Poisson distribution as a limit of the Bernoulli counting process, if we view N_t as a a stochastic process $(N_t)_{t \geq 0}$ (such that N_t counts the total number of arrivals by time t), it will inherit the properties of the counting process.

Theorem 3.2

Let $(N_t)_{t\geq 0}$ be a Poisson distribution. Then we have

- $(N_t)_{t\geq 0}$ is stationary. That is, for any t and any s, $N_{t+s} - N_s \sim N_t \sim \text{Poisson}(\lambda t)$.
- $(N_t)_{t\geq 0}$ has independent increments. That is, $N_t - N_s$ is independent from $N_u - N_v$, as long as the intervals $[s,t]$ and $[v,u]$ are disjoint.

Theorem 3.3

Let $X \sim \text{Poisson}(\lambda), Y \sim \text{Poisson}(\mu)$, with X independent of Y. Then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

Proof. Let $(N_t)_{t\geq 0}$ be a Poisson process of parameter 1, such that $N_s \sim \text{Poisson}(s)$. Moreover, we have $N_t - N_s \sim \text{Poisson}(t-s)$ independent of N_s. Then we can insert X and Y into this process by noting that $X \sim N_\lambda$, and $Y \sim N_{\lambda+\mu} - N_\lambda$, with independence preserved. So $X + Y \sim N_\lambda + N_{\lambda+\mu} - N_\lambda = N_{\lambda+\mu}$. So $X + Y \sim \text{Poisson}(\lambda + \mu)$. □

Definition 3.4

Given a random variable X, define its **cumulative distribution function** as

$$F_X(x) \coloneqq \mathbb{P}(X \leq x)$$

If X is continuous, then this is

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

Theorem 3.4

Let X be a continuous random variable. Then $f_X(x) = \frac{d}{dx}F_X(x)$.

Suppose we ask the time that the kth arrival occurs, $\tau_k \coloneqq \inf\{t > 0 | N_t = k\}$. Then we have

$$\mathbb{P}(\tau_1 > t) = \mathbb{P}(N_t = 0) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}|_{k=0} = e^{-\lambda t}$$

Then if we consider the cumulative distribution function of τ_1, we see that by definition,

$$F_{\tau_1}(t) \coloneqq \mathbb{P}(\tau_1 \leq t) = 1 - \mathbb{P}(\tau_1 > t) = 1 - e^{-\lambda t}$$

Then to find the probability density function, we simply differentiate the cumulative density function

$$f_{\tau_1}(t) = F'_{\tau_1}(t) = \frac{d}{dt}[1 - e^{-\lambda t}] = \lambda e^{-\lambda t}$$

So we have found the distribution of the first arrival time:

$$f_{\tau_1}(t) = \lambda e^{-\lambda t} 1_{[0,\infty)}(t)$$

Definition 3.5

Suppose a random variable τ has a probability density function $f_\tau(t) = \lambda e^{-\lambda t} 1_{[0,\infty)}(t)$. Then we say that $\tau \sim \text{Exponential}(\lambda)$.

So we have seen that $\tau_1 \sim \text{Exponential}\,\lambda$. In the general case, we can ask for the distribution of τ_k. Then we have

$$\mathbb{P}(\tau_k > t) = \mathbb{P}(N_t < k) = \sum_{i=0}^{k-1} \mathbb{P}(N_t = i) = \sum_{i=0}^{k-1} \frac{(\lambda t)^i}{i!} e^{-\lambda t}$$

Using the cumulative density function, we have

$$f_{\tau_k}(t) = F'_{\tau_k}(t) = \frac{d}{dt}[1 - \mathbb{P}(\tau_k > t)] = \frac{d}{dt}[e^{-\lambda t} \sum_{i=0}^{k-1} \frac{(\lambda t)^i}{i!}]$$

$$-\lambda e^{-\lambda t} \sum_{i=0}^{k-1} \frac{(\lambda t)^i}{i!} + e^{-\lambda t} \lambda \sum_{i=1}^{k-1} \frac{(\lambda t)^{i-1}}{(i-1)!} = -\lambda e^{-\lambda t} [\sum_{i=0}^{k-2} \frac{(\lambda t)^i}{i!} - \sum_{i=0}^{k-1} \frac{(\lambda t)^i}{i!}]$$

$$= e^{-\lambda t} \lambda^k \frac{t^{k-1}}{(k-1)!} 1_{[0,\infty)}(t)$$

Definition 3.6

If a random variable X has probability density function

$$f_X(x) = \frac{x^{k-1}}{(k-1)!} \lambda^k e^{-\lambda t} 1_{[0,\infty)}(x)$$

for some $k \geq 1$ and $\lambda > 1$, then we say X has the **gamma distribution** with shape k and rate λ, written $X \sim \Gamma(k, \lambda)$.

Then we can see that for a Poisson distribution with parameter λ, we have $\tau_k \sim \Gamma(k, \lambda)$.

Example 3.1

Suppose we have $X \sim \text{Exponential}(\lambda)$ for some $\lambda > 0$. What is $\mathbb{E}[X^k]$ for $k \geq \mathbb{N}$?

Let $g(X) = X^k$. Then we want to find $\mathbb{E}[g(X)]$. By the Law of the Unconscious Statistician, we have

$$\mathbb{E}[g(X)] = \int_0^\infty g(x) f_X(x) dx$$

Consider the case where $k = 0$, so $g = 1$. Then we have

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

Differentiate on both sides with respect to λ (derivative and integral can be interchanged):

$$\int_0^\infty -xe^{-\lambda x} dx = -\frac{1}{\lambda^2}$$
$$\implies \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

But by the Law of the Unconscious Statistician, the left hand side is precisely

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

Of course, this makes sense since $\mathbb{E}[X]$ is precisely the expected value of the first arrival time of a Poisson process. For higher k, we simply keep differentiating with respect to λ (implicitly cancelling negative signs on both sides):

$$\int_0^\infty xe^{-\lambda x} dx = \frac{1}{\lambda^2}$$
$$\implies \int_0^\infty x^2 e^{-\lambda x} dx = \frac{2}{\lambda^3}$$
$$\implies \int_0^\infty x^3 e^{-\lambda x} dx = \frac{2 * 3}{\lambda^4}$$
$$\implies \dots$$

So we find that in general,

$$\int_0^\infty x^k e^{-\lambda x} dx = \frac{k!}{\lambda^{k+1}}$$

Then we can move a λ to the left side and solve for $\mathbb{E}[X^k]$:

$$\mathbb{E}[X^k] = \int_0^\infty x^k \lambda e^{-\lambda x} dx = \frac{k!}{\lambda^k}$$

Example 3.2

Suppose children are born in a certain hospital at a rate of $\lambda = 5$ each day. What is the probability at least 2 are born in the next 6 hours?

We are interested in $\mathbb{P}(N_{1/4} \geq 2)$. Then this is

$$\mathbb{P}(N_{1/4} \geq 2) = 1 - \mathbb{P}(N_{1/4} = 0) - \mathbb{P}(N_{1/4} = 1) = 1 - \frac{(\frac{5}{4})^0 e^{-5/4}}{0!} - \frac{(\frac{5}{4})^1 e^{-5/4}}{1!}$$
$$= 1 - e^{-5/4} - \frac{5}{4} e^{-5/4}$$

Example 3.3

Suppose you catch fish at a rate of $\lambda = 2$ per hour, starting at 10 am. What is the probability you catch exactly 1 by 10:30, and exactly 3 total by 12?

We want $\mathbb{P}(N_{1/2} = 1, N_2 = 3)$. We can split the second term to get

$$\mathbb{P}(N_{1/2} = 1, N_2 - N_{1/2} + N_{1/2} = 3) = \mathbb{P}(N_{1/2} = 1, N_2 - N_{1/2} = 3 - N_{1/2})$$

Since we intersect only with events where $N_{1/2} = 1$, we can plug this into the right term:

$$\mathbb{P}(N_{1/2} = 1, N_2 - N_{1/2} = 3 - N_{1/2}) = \mathbb{P}(N_{1/2} = 1, N_2 - N_{1/2} = 2)$$

Since $N_{1/2}$ and $N_2 - N - 1/2$ are independent, we have

$$\mathbb{P}(N_{1/2} = 1, N_2 - N_{1/2} = 2) = \mathbb{P}(N_{1/2} = 1)\mathbb{P}(N_2 - N_{1/2} = 2)$$

By stationarity, this is

$$\mathbb{P}(N_{1/2} = 1)\mathbb{P}(N_2 - N_{1/2} = 2) = \mathbb{P}(N_{1/2} = 1)\mathbb{P}(N_{3/2} = 2)$$

Then we can plug into the probability density function to find:

$$\mathbb{P}(N_{1/2} = 1, N_{3/2} = 2) = \mathbb{P}(N_{1/2} = 1)\mathbb{P}(N_{3/2} = 2) = \frac{(2 * \frac{1}{2})^1 e^{-2*1/2}}{1!} * \frac{(2 * \frac{3}{2})^2 e^{2*3/2}}{2!}$$

3.3 Superposition and Thinning

Recall that if we say that $(N_t)_{t \geq 0}$ is a Poisson process with parameter λ counting the number of arrivals by time t, then it satisfies three properties:

- $N_0 = 0$: the count begins at 0
- $N_t - N_s \sim N_{t-s} \sim \text{Poisson}(\lambda(t - s))$: stationarity

Bernoulli Processes	Poisson Processes
$S_{n+m} - S_m \sim S_n \sim \text{Binom}(n,p)$	$N_{t+s} - N_s \sim N_t \sim \text{Poisson}(\lambda t)$
$S_n - S_m$ independent $S_k - S_p$ $0 \le m \le n \le p \le k$	$N_t - N_s$ independent $N_u - N_v$ $0 \le s \le t \le v \le u$
$f_{T_k}(n) = (1-p)^{n-1}p \sim \text{Geom}(p)$	$f_{\tau_1}(t) = \lambda e^{-\lambda t} 1_{[0,\infty)}(t) \sim \text{Exponential}(\lambda)$
$\mathbb{E}[T_1] = \frac{1}{p}$	$\mathbb{E}[\tau_1] = \frac{1}{\lambda}$
$f_{T_k}(n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$	$f_{\tau_k}(t) = \frac{t^{k-1}}{(k-1)!} \lambda^k e^{-\lambda t} 1_{[0,\infty)}(t)$
$\mathbb{E}[T_k] = \frac{k}{p}$	$\mathbb{E}[\tau_k] = \frac{k}{\lambda}$

- $(s,t] \cap (u,v] = \varnothing \implies N_t - N_s$ and $N_v - N_u$ independent: Independent increments

We want to investigate a property of Poisson processes known as the *Markov property*. Informally, this property means that the process can "start fresh" independent of past activity.

Example 3.4

Fix some $r > 0$. Define $\tilde{N}_t \coloneqq N_{r+t} - N_r$. Then we claim that $(\tilde{N}_t)_{t \ge 0}$ is also a Poisson process with rate λ. To do this, let us check the three properties:

- $\tilde{N}_0 = N_{r+0} - N_r = N_r - N_r = 0$
- $\tilde{N}_t - \tilde{N}_s = N_{r+t} - N_{r+s} \sim N_{t-s} \sim \text{Poisson}(\lambda(t-s))$
- If we have $\tilde{N}_v - \tilde{N}_u$ and $\tilde{N}_t - \tilde{N}_s$ with $(s,t] \cap (u,v] = \varnothing$, then $(r+s, r+t] \cap (r+u, r+v] = \varnothing$ so independent increments holds from the original.

Moreover, we can see from independent increments that $(\tilde{N}_t)_{t \ge 0}$ is independent of the values of N_t for $0 \le t \le r$.

Theorem 3.5: Superposition

Let $(N_t^1)_{t \ge 0}, \ldots, (N_t^k)_{t \ge 0}$ be Poisson processes of respective rates $\lambda_1, \ldots, \lambda_k > 0$. If the processes are all independent, then the process $(N_t)_{t \ge 0}$ with $N_t \coloneqq N_t^1 + \ldots + N_t^k$ is another Poisson Process of rate $\lambda = \lambda_1 + \ldots + \lambda_k$.

This combining process is known as superposition. If we imagine marking the arrival times of each individual process on a number line, then the arrival times of the combined process is simply the superimposed image of all the individual number lines.

In the opposite direction, we can also apply a process called thinning, which will separate the arrivals of one process into different categories.

Example 3.5

Suppose $(N_t)_{t\geq 0}$ is a Poisson process with parameter $\lambda > 0$. For each arrival τ_k, flip a coin and record the result. Define the process $(N_t^1)_{t\geq 0}$ which counts the number of arrivals by time t where a heads was flipped. Similarly, define $(N_t^2)_{t\geq 0}$ to count the number of arrivals where a tails was flipped.

In this manner, we have thinned our original process into two subprocesses. Note that if we superimposed the thinned processes, we would reconstruct the original process.

Theorem 3.6: Thinning

Suppose $(N_t)_{t\geq 0}$ is a Poisson process with rate $\lambda > 0$. Suppose each arrival is assigned, independent of (N_t) and independent of each other, one of k types with respective probabilities $p_1, \ldots, p_k$, with $p_1 + \ldots + p_k = 1$. Define N_t^i to be the number of arrivals of type i by time t, with $1 \leq i \leq k$. Then the $(N_t^i)_{t\geq 0}$ are independent Poisson processes, which have respective rates $p_1\lambda, \ldots, p_k\lambda$.

Example 3.6

Suppose passengers arrive at a bus stop at rate $\lambda > 0$, and buses arrive at rate $\mu > 0$. Let N_t^1 and N_t^2 denote the Poisson processes for the passengers and buses, respectively. What is the probability that k passengers get on the first bus?

While we could calculate this directly, consider the superimposed process N_t, which counts arrivals of both buses and people. Then the probability that k passengers get on the bus is the probability that the first k arrivals in N_t are passengers, and the $k+1$th arrival is a bus. By superposition, N_t is a Poisson process of rate $\lambda + \mu$. Moreover, we showed in homework that the probability that the first arrival is a passenger is $\frac{\lambda}{\lambda+\mu}$, and the probability that it is a bus is $\frac{\mu}{\lambda+\mu}$. Then the probability is

$$\left(\frac{\lambda}{\lambda+\mu}\right)^k \frac{\mu}{\lambda+\mu}$$

Using the logic from this example, we have the following

Corollary

Suppose $\sigma^{(1)}, \ldots, \sigma^{(k)}$ are independent exponential variables with respective rates $\lambda_1, \ldots, \lambda_k > 0$. Then

- $\sigma \coloneqq \min\{\sigma^{(1)}, \ldots, \sigma^{(k)}\} \sim \text{Exponential}(\lambda_1 + \ldots + \lambda_k) = \text{Exponential}(\lambda)$
- $\mathbb{P}(\sigma = \sigma^{(i)}) = \lambda_i/\lambda$

In other words, the first arrival time of any of the respective Poisson processes is exponential with rate equal to the sum of the original rates, and the probability that that arrival is from process i is simply the proportion of the total rate which is contributed by λ_i.

Proof. The first point is essentially proved by the example above.

If we let N_t^i be the Poisson process associated with $\sigma^{(i)}$, then each N_t^i is a thinning of some N_t, with a thinning rate of some p_i for each i. Then if we let λ be the rate of this N_t, then we must have $\lambda_i = p_i\lambda$. So we have $p_i = \lambda_i/\lambda$. □

Example 3.7

Suppose Y is a sum of N random variables $X_1, \ldots, X_N$, where the X_i are i.i.d. as Exponential(λ), and $N \sim \text{Geom}(p)$. What is the distribution of Y?

Let N_t be a Poisson process of rate $\lambda > 0$. Note that we can then think of each X_k as the interarrival time $\tau_k - \tau_{k-1}$. Then attach to each arrival a label (independent of each other and of N_t)) of either L_1 or L_2, where the probability of L_1 is p. Then we can think of Y as the first arrival time of type L_1. Specifically, let N_t^1 be the process N_t, thinned out at rate p. Then Y is the first arrival time of N_t^1, and thus we have $Y \sim \text{Exponential}(p\lambda)$.

Chapter 4

Limit Theorems

4.1 Variance and Covariance

Recall from the frequentist approach that one definition of the probability of an event is calculated by conducting arbitrarily many trials and observing the ratio as the number goes to infinity:

$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{S_N}{N}$$

Here, the left side of the equation is just a number, so there is no randomness, but the inside of the limit is a random variable that depends on the specific experiment outcomes. So we observe that the "randomness" of the quantity $\frac{N_A}{N}$ must go to zero as our experiments increase. This motivates a way to quantify randomness:

Definition 4.1

Given a random variable X, the **variance** of X is

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

From a simple observation, for a constant variable X, $X = \mathbb{E}[X]$ and thus $\mathrm{Var}(X) = 0$. On the other hand, for a variable with high variance, X is often very far from $\mathbb{E}[X]$. Thus, the variance tells us how good of an approximation $\mathbb{E}[X]$ is for X.

A more general quantity that we can compute is the covariance:

Definition 4.2

Given two random variables, X, Y, the **covariance** of X and Y is

$$\mathrm{Cov}(X, Y) = \sigma_{X,Y} \coloneqq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

This essentially tells us how correlated two variables are. We call two variables *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$. In particular, if we have independent variables X, Y, then $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[X - \mathbb{E}[X]]\mathbb{E}[Y - \mathbb{E}[Y]] = 0$. So independent variables are

uncorrelated (but the converse is not necessarily true). Moreover, $\operatorname{Var}(X) = \operatorname{Cov}(X, X)$.

The covariance has the important property of being *bilinear*. That is, it is linear in each of the arguments independently, such that

$$\operatorname{Cov}(aX + Y, Z) = a \operatorname{Cov}(X, Z) + \operatorname{Cov}(Y, Z)$$
$$\operatorname{Cov}(X, aY + Z) = a \operatorname{Cov}(X, Y) + \operatorname{Cov}(X, Z)$$

If we modify both arguments at once, we have:

$$\operatorname{Cov}(aW + bX, cY + dZ) = ac \operatorname{Cov}(W, Y) + ad \operatorname{Cov}(W, Z) + bc \operatorname{Cov}(X, Y) + bd \operatorname{Cov}(X, Z)$$

More generally, we have the following:

Proposition 4.1

Given any scalars $a_i, b_j \in \mathbb{R}$ and random variables X_i, Y_j, we have

$$\operatorname{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \operatorname{Cov}(X_i, Y_j)$$

This allows us to easily prove a number of nice properties about variance and covariance:

Corollary

For any random variables X, Y, we have:

1. $\operatorname{Var}(aX + b) = a^2 \operatorname{Var} X$ for any $a, b \in \mathbb{R}$.
2. $\operatorname{Var}(X) \geq 0$.
3. $\operatorname{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
4. $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2 \operatorname{Cov}(X, Y)$.
5. If X, Y are independent, then $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.

Proof. 1) Note that the addition of b doesn't change anything, since the quantity $X - \mathbb{E}[X]$ is the same regardless. Then we have

$$\operatorname{Var}(aX) = \operatorname{Cov}(aX, aX) = a^2 \operatorname{Cov}(X, X) = a^2 \operatorname{Var}(X)$$

2) Since the quantity $(X - \mathbb{E}[X])^2 \geq 0$, we have

$$\operatorname{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \geq 0$$

3) By linearity,

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

Since $\mathbb{E}[X]$ is a constant,

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

4) Again, we can expand with linearity:

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y-\mathbb{E}[X+Y])^2] \\
&= \mathbb{E}[(X+Y-\mathbb{E}[X]-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[((X-\mathbb{E}[X])+(Y-\mathbb{E}[Y]))^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] + 2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)
\end{aligned}$$

5) Follows from 4) since $\mathrm{Cov}(X,Y) = 0$ when X, Y are independent. □

Example 4.1

Suppose $\tau \sim \exp(\lambda)$. Then we have $\mathbb{E}[\tau] = \frac{1}{\lambda}$. As we have previously calculated,

$$\mathbb{E}[\tau^k] = \frac{k!}{\lambda^k}$$

So we have

$$\mathrm{Var}(\tau) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Then if we have $\gamma \sim \Gamma(k, \lambda)$, then $\gamma \sim \tau_1 + \ldots + \tau_k$ and thus

$$\mathrm{Var}(\gamma) = \mathrm{Var}(\tau_1 + \ldots + \tau_k) = k\,\mathrm{Var}(\tau_1) = \frac{k}{\lambda^2}$$

4.2 The Law of Large Numbers

Now let us revisit the situation we began this chapter with. If we conduct N trials of an experiment, and let S_N be the number of times the event A occurs, then the frequentist approach claims that we should define

$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{S_N}{N}$$

As we observed earlier, for this equation to make sense then the quantity $\frac{S_N}{N}$ must converge to a consistent quantity; that is, the variance should be zero. To see this, observe that

$$\mathrm{Var}(\frac{S_N}{N}) = \frac{1}{N^2}\,\mathrm{Var}(S_N) = \frac{1}{N^2} N\,\mathrm{Var}(1_A) = \frac{\mathrm{Var}(1_A)}{N}$$

Since $\mathrm{Var}(1_A)$ is a constant, we have

$$\lim_{N\to\infty} \mathrm{Var}\left(\frac{S_N}{N}\right) = \lim_{N\to\infty} \frac{\mathrm{Var}(1_A)}{N} = 0$$

This result is known as the Law of Large Numbers. This law states informally that although S_N/N may deviate from $\mathbb{P}(A)$ for finite N, it will converge to a consistent limit, and moreover that $\mathbb{P}(A)$ will be that limit. In other words, the distribution of S_N/N will accumulate around $\mathbb{P}(A)$ as $N \to \infty$. There are two formulations of this. The first is what we have just proved:

Theorem 4.2: Weak Law of Large Numbers

Let X be a random variable with $\mathbb{E}[X^2] < \infty$. Let $X_1, X_2, \ldots$ be i.i.d according to X. Define $S_n = \sum_{i=1}^n X_i$. Then the quantity S_n/n converges in L^2 to $\mathbb{E}[X]$, that is,

$$\mathbb{E}\left[\left(\frac{S_n}{n} - \mathbb{E}[X]\right)^2\right] \to 0$$

Proof. Since $\mathbb{E}[S_n/n] = n\mathbb{E}[X/n] = \mathbb{E}[X]$, this quantity is just the variance of S_n/n. From the previous discussion, we then have

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{S_n}{n} - \mathbb{E}[X]\right)^2\right] &= \lim_{n\to\infty} \mathrm{Var}\left(\frac{S_n}{n}\right) = \lim_{n\to\infty} \frac{\mathrm{Var}(S_n)}{n^2} \\ &= \lim_{n\to\infty} \frac{\mathrm{Var}(X_1 + \ldots + X_n)}{n^2} = \lim_{n\to\infty} \frac{\mathrm{Var}(X)}{n} = 0\end{aligned}$$

□

There is also a stronger formulation of this statement, known as the Strong Law of Large Numbers. This formulation says that S_n/n converges to $\mathbb{E}[X]$ almost surely; that is, for nearly all $\omega \in \Omega$ (or on a set of measure 1).

Theorem 4.3

Let X be a random variable with $\mathbb{E}[X^4] < \infty$. Let $X_1, X_2, \ldots$ be i.i.d. according to X. Define $S - n = \sum_{i=1}^n X_i$. Then $\mathbb{P}(S_n/n \to \mathbb{E}[X]) = 1$.

4.3 The Central Limit Theorem

The Law of Large Numbers tells us that as we take repeated values of a random variable, the average value will converge to a single mean value (as long as $\mathbb{E}[X^2] < \infty$, and almost surely when $\mathbb{E}[X^4] < \infty$). If we simulate for large (finite) values of n, we find that the distribution of S_n/n not only centers about $\mathbb{E}[X]$, but tends to do so in the shape of a "bell curve." In fact, this shape appears at relatively low values of n (even 10 or 20 trials is enough to begin discerning a bell curve). Of course, in the limit, this bell curve becomes infinitely thin and is the Dirac delta function.

For instance, if we roll a die many times, even though each individual value is equally likely, it is unlikely that the average will be near 1, since this requires rolling lots of low numbers. Similarly, the average is unlikely to be near 6. On the other hand, it is far more likely that the average value is near 3.5, the expectation of one roll.

This makes intuitive sense, since extreme behavior after repetition requires many instances of extreme behavior *in the same direction*, when it is more likely that some of the extreme behavior will act in the opposite direction and cancel out opposing behavior. Thus, in the limit, we are less likely to get extreme behavior than to get behavior near the average. This observation is formalized by the Central Limit Theorem.

Definition 4.3

A random variable X has the **normal distribution** with mean μ and variance σ^2, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if it is a continuous variable with probability density function

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

In particular, the *standard normal* is $\mathcal{N}(0, 1)$.

Changing the parameters of the normal distribution amounts to shifting or scaling it, but the shape stays the same regardless. In order to describe this shape without regard to the numerical value of the parameters, we will *standardize* the random variable.

Definition 4.4

The **standard deviation** of a random variable X is $\sigma := \sqrt{\mathrm{Var}(X)}$.

Suppose we have some random variable Y. Then to standardize it, we would like to shift this variable to $Y' = Y'(Y)$ so that the following properties hold: $\mu_{Y'} = 0$ and $\sigma^2_{Y'} = \sigma_{Y'} = 1$. Suppose we do this by setting

$$Y'(Y) := \frac{Y - \mu_Y}{\sigma_Y}$$

Then we clearly have $\mu_Y = 0$. To check the variance, note that

$$\mathrm{Var}(Y') = \frac{1}{\sigma_Y^2}\mathrm{Var}(Y - \mu_Y) = \frac{1}{\sigma_Y^2}\mathrm{Var}(Y) = 1$$

So given any random variable, we can standardize it so that it has mean 0 and variance 1. When we do this for a normally distributed variable, we often call this a *z-score*. We will apply a technique commonly used in statistics, where we identify the location of a given point relative to the distribution.

Definition 4.5

Given a normal distribution $\mathcal{N}(\mu, \sigma^2)$ and some value $x \in \mathbb{R}$, the **z-score** of x with respect to $\mathcal{N}(\mu, \sigma^2)$ is

$$z := \frac{x - \mu}{\sigma}$$

In other words, rather than identifying a given data point with its actual value, we will simply measure how many standard deviations it is above or below the mean.

We are now prepared to state and prove the Central Limit Theorem:

Theorem 4.4: Central Limit Theorem

Let $X_1, X_2, \dots$ be i.i.d according to a random variable X with $\mu = \mathbb{E}[X] \in \mathbb{R}$ and $0 < \sigma^2 = \text{Var}(X) < \infty$. Let $S_n \coloneqq \sum_{i=1}^n X_i$ be the sum of the first n values of X_i. Let $\overline{X}_n \coloneqq S_n/n$ be the average of the first n values of X_i.

Let $\mu_n = \mu_{\overline{X}_n}$ and $\sigma_n = \sigma_{\overline{X}_n}$ be the mean and standard deviations of $\overline{X}_n$, respectively. Let Z_n be the standardization of $\overline{X}_n$, that is,

$$Z_n \coloneqq \frac{\overline{X}_n - \mu_n}{\sigma_n} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

Then as $n \to \infty$, the distribution of Z_n tends to $\mathcal{N}(0,1)$.

Another way of viewing this is by looking at relative cumulative densities. That is, if we let Y be a placeholder variable with $Y \sim \mathcal{N}(0, \sigma^2)$, then

$$\lim_{n\to\infty} \mathbb{P}(a < \sqrt{n}(\overline{X} - \mu) \le b) = \mathbb{P}(a < Y \le b) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$

In general, this may not be easy to calculate, as the integral of $\exp(-x^2)$ is not able to be expressed in terms of elementary functions. However, we can often use computers or tables to approximate the values. We can also use DeMoivre's rule, or the 68-95-99.7 rule, to provide simple approximations. This rule states that for a normal distribution, 68% of the data falls within one standard deviation of the mean, 95% within two, and 99.7% within 3.

In order to simply calculations, we will sometimes adopt two conventions for the probability density function and cumulative distribution functions of the normal distribution. Given some $X \sim \mathcal{N}(\mu, \sigma^2)$, we write $\varphi_X(x) \coloneqq f_X(x)$, and $\Phi_X(x) \coloneqq F_X(x) = \int_{-\infty}^x \varphi_X(x)dx$.

Example 4.2

Let X be the indicator function of some event A, so $X = 1_A$. Let $X_1, \dots$ be i.i.d. trials of A, so $X_i = 1_{A_i}$. Suppose $\mathbb{E}[X] = \mathbb{P}(A) = p$. Then

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

Now in the case of an indicator function, $X = 1$ or 0, so $X^2 = X$ and we can say

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p)$$

Then letting S_n be the sum of the first n indicator functions and Z_n be the standardization, then the central limit theorem says that

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - np}{\sqrt{n}\sqrt{p(1-p)}} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

Now consider a *simple random walk*. That is, suppose we start at position 0, and take steps $\xi_1, \xi_2, \ldots$, with $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = 1/2$. Then we can rewrite this as $\xi_i = 2X_i - 1$, where $X_i \sim \text{Bernoulli}(p)$. So $\mu_\xi = 0$ and $\text{Var}(\xi) = \text{Var}(2X - 1) = 4\,\text{Var}(X) = 4p(1-p)$. Since we have $p = 1/2$, this is $\text{Var}(\xi) = 1$.

4.4 Brownian Motion

We can use the limit theorems that we have just discussed to derive another example of a stochastic process, known as Brownian motion. This process arose from observations in biology, where nonliving particles were noted to spontaneously move in a chaotic pattern at the microscopic level. Einstein later theorized that this motion arose through impacts with randomly moving individual water molecules. Thus, this chaotic motion was the sum of small impacts, each in a random direction.

If we let these impacts be represented by $\xi_1, \xi_2, \ldots$, then we say that

$$B_t^{(N)} := \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \xi_K = \sqrt{N}\,\overline{\xi}_{Nt} \quad \overset{N\to\infty}{\longrightarrow} \quad B_t$$

Here, B_t is an example of *Brownian motion*.

Definition 4.6

A stochastic process $(B_t)_{t\geq 0}$ is said to be a **standard Brownian Motion** in one dimension if:

- $B_0 = 0$.
- The function $t \mapsto B_t$ is continuous.
- $B_t - B_s \sim B_{t-s} \sim \mathcal{N}(0, t-s) \sim \sqrt{t-s}B_1$.
- $B_t - B_s$ is independent of $B_v - B_u$ if $(s,t] \cap (u,v] = \varnothing$.

It follows that for any time t, $B_t \sim \mathcal{N}(0,t)$.

Intuitively that Brownian motion is a continuous time analogue of the simple random walk, where the steps are taken to be infinitely small but happening infinitely often.

Now that we have derived Brownian motion, we can use it to prove properties of Gaussian distributions, similarly to how we used the Poisson process to prove properties of Poisson distributions.

Proposition 4.5

Here are some properties of Gaussian distributions:

1. If $X \sim \mathcal{N}(0, \sigma^2) \perp Y \sim \mathcal{N}(0, \eta^2)$, then $X + Y \sim \mathcal{N}(0, \sigma^2 + \eta^2)$.
2. If $Z \sim \mathcal{N}(0, 1)$, then $\sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$.
3. If $X \sim \mathcal{N}(0, a)$, then $X \sim \sqrt{a}\mathcal{N}(0, 1) \sim \sqrt{a}Z$.
4. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the z-score $Z \coloneqq \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
5. $X_1, \ldots, X_n$ are i.i.d with distribution $\mathcal{N}(0, 1)$, then $\sum a_i X_i \sim \mathcal{N}(0, \sum a_i^2)$.

Proof. 1. If we consider a standard Brownian process $(B_t)_{t\geq 0}$, then $X \sim B_{\sigma^2}$ and $Y \sim B_{\sigma^2+\eta^2} - B_{\sigma^2}$ (by stationarity). By independent increments,

$$X + Y \sim B_{\sigma^2} + B_{\sigma^2+\eta^2} - B_{\sigma^2} = B_{\sigma^2+\eta^2} \sim \mathcal{N}(0, \sigma^2 + \eta^2)$$

2. Using dx notation, we have

$$f_{\sigma Z+\mu}(x)dx = \mathbb{P}(\sigma Z + \mu \in [x, x + dx])$$
$$= \mathbb{P}(\sigma Z \in [x - \mu, x - \mu + dx]) = \mathbb{P}(Z \in [\underbrace{\frac{x-\mu}{\sigma}}_{z}, \underbrace{\frac{x-\mu}{\sigma}}_{z} + \underbrace{\frac{dx}{\sigma}}_{dz}])$$
$$= f_Z(z)dz = \frac{1}{\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{\sigma^2}} \frac{dx}{\sigma}$$

which is simply the pdf of a function with distribution $\mathcal{N}(\mu, \sigma^2)$. □

The most important results of the above proposition can be neatly summarized using the following important fact: **linear functions of independent Gaussians are Gaussian**. If we remove the independence assumption, however, this may not be true.

When we combine Gaussian distributions, we can either combine them to create a *multivariate* Gaussian or a *joint* Gaussian.

Definition 4.7

A random vector $\overrightarrow{Z} \coloneqq (X, Y)$ is a **multivariate Gaussian distribution** with mean $\overrightarrow{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$ and *covariance matrix* $\Sigma = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{bmatrix}$ if the distribution of Z (or the joint distribution of (X, Y) is given by

$$f_Z(z) = f_{(X,Y)}(x, y) = \frac{\exp(\frac{1}{2}(\overrightarrow{z} - \overrightarrow{\mu})^T \Sigma^{-1} (\overrightarrow{z} - \overrightarrow{\mu}))}{\sqrt{(2\pi)^2 \det(\Sigma)}}$$

Here, the covariance matrix takes the place of the variance for the typical distribution. This definition can easily be extended to any finite number of variables $X_1, \ldots, X_n$, simply

by increasing the size of the vectors and matrices, and increasing the power of 2π.

Note that we suggestively labelled the above distribution a Gaussian distribution, but we did not actually demand that the X_i were Gaussian variables themselves. However, we can prove that this must be the case, and in fact we have a slightly stronger condition.

Definition 4.8

n variables $X_1, \dots, X_n$ are **jointly Gaussian** if for any $a_1, \dots, a_n$, the linear combination $a_1 X_+ \dots + a_n X_n$ is a (univariate) Gaussian distribution.

Note that if we already know the X_i are Gaussian, then this is a weaker condition than independence by the previous proposition. On the other hand, if we know the X_i are jointly Gaussian, then by setting all of the a_i to 0 except for 1, we see that each of the X_i must be Gaussian on their own.

Theorem 4.6

$\overrightarrow{Z} = (X_1, \dots, X_n)$ is multivariate Gaussian if and only if the $X_1, \dots, X_n$ are jointly Gaussian.

If we graph a simple random walk, we will see that it consists of jagged triangular paths. In the limit, then, Brownian motion is similarly jagged, but it is jagged everywhere. Similar to the Weierstrass function, we have the following fact:

Proposition 4.7

The function $t \mapsto B_t$ is nowhere differentiable (with probability 1).

Proof. We have $\frac{d}{dt} B_t \coloneqq \lim_{\varepsilon \to 0} \frac{B_{t+\varepsilon} - B_t}{\varepsilon} \sim \lim_{\varepsilon \to 0} \frac{B_\varepsilon}{\varepsilon} \sim \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathcal{N}(0, \varepsilon) \sim \lim_{\varepsilon \to 0} \mathcal{N}(0, \frac{1}{\varepsilon})$. This limit doesn't exist, so the derivative does not exist (with probability 1). □

Analogously to the the above facts about Gaussians, we have the following:

Proposition 4.8

If $(B_t)_{t \geq 0}$ is Brownian motion, then $(-B_t)_{t \geq 0}$ and $(\tilde{B}_s)_{s \geq 0} \coloneqq (B_{t+s} - B_t)_{t \geq 0}$ are both Brownian motion.

Theorem 4.9: Reflection Principle

Let $(B_t)_{t\geq 0}$ be Brownian motion. For any number $b > 0$, define the arrival time at b to be $\tau_b := \min\{t : B_t = b\}$. Then the probabilities that (B_t) ends up above b is

$$\mathbb{P}(B_t > b) = \frac{1}{2}\mathbb{P}(\tau_b \leq t)$$

That is, for every path which ends up "above b" (at time t), there are twice as many paths which touch b (by time t).

Proof. By the law of total probability, we have

$$\mathbb{P}(B_t > b) = \mathbb{P}(B_t > b|\tau_b \leq t)\mathbb{P}(\tau_b \leq t) + \underbrace{\mathbb{P}(B_t > b|\tau_b > t)}_{=0}\mathbb{P}(\tau_b > t) = \frac{1}{2}\mathbb{P}(\tau_b \leq t)$$

The term $\mathbb{P}(B_t > b|\tau_b \leq t) = \frac{1}{2}$ because of the symmetry of Brownian motion. For any path starting at $B_{\tau_b} = b$ with $B_t > b$, the inverted path which also starts at $B_{\tau_b} = b$ has $B_t < b$. So only half the paths starting at $B_{\tau_b} = b$ have $B_t > b$. □

This proof is helpful because it allows us to more easily calculate the pdf of τ_b. By definition, we have

$$F_{\tau_b}(t) = \mathbb{P}(\tau_b \leq t)$$

By the reflection principle, this is

$$\mathbb{P}(\tau_b \leq t) = 2\mathbb{P}(B_t > b) = 2\mathbb{P}(\sqrt{t}B_1 > b) = 2\mathbb{P}(B_1 > \frac{b}{\sqrt{t}}) = 2\int_{\frac{b}{\sqrt{t}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

And using the fundamental theorem of calculus:

$$f_{\tau_b}(t) = \frac{d}{dt}F_{\tau_b}(t) = \frac{d}{dt}2\int_{\frac{b}{\sqrt{t}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = -\frac{2}{\sqrt{2\pi}} e^{-b^2/2t}(-\frac{1}{2}\frac{b}{t^{3/2}}) = \frac{1}{\sqrt{2\pi}} e^{-b^2/2t}\frac{b}{2t^{3/2}}$$

This allows us to easily extend to negative values of b:

$$f_{\tau_b}(t) = \frac{|b|e^{-b^2/2t}}{\sqrt{2\pi}t^{3/2}}$$

Then we can calculate the probability that a standard Brownian motion ever hits a certain threshold value a:

$$\mathbb{P}(\tau_a < \infty) = \lim_{t\to\infty} F_{\tau_a}(t) = 2\int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} = 1$$

Thus, Brownian motion hits every threshold value with probability 1. However, the time it takes to do that is infinite in expectation:

$$\mathbb{E}[\tau_a] = \int_0^{\infty} t\frac{ae^{-a^2/2t}}{\sqrt{2\pi}t^{3/2}}dt = \int_0^{\infty} \frac{ae^{-a^2/2t}}{\sqrt{2\pi t}}dt$$

However, this integral diverges, since $e^{-a^2/2t} \approx 1$ for large t, and the integral $\int_b^\infty \frac{1}{\sqrt{t}}$ diverges. So $\mathbb{E}[\tau_a] = \infty$.

Moreover, we can ask about the probability that we end up above a at any point in an interval $[0, t]$:

$$\mathbb{P}((\max_{0 \le s \le t} B_s) \ge a) = \mathbb{P}(\tau_a \le t) = 2\mathbb{P}(B_t \ge a)$$

by the reflection principle. But we also have the following:

$$\mathbb{P}(|B_t| \ge a) = \mathbb{P}(B_t \ge a \text{ or } B_t \le -a) = \mathbb{P}(B_t \ge a) + \mathbb{P}(-B_t \ge a) = 2\mathbb{P}(B_t \ge a)$$

In particular, we have the following fact:

Theorem 4.10

If $|B_t|$ is the absolute value of standard Brownian motion, $|B_t|$, and we let $\max_{s\in[0,t]} B_s$ be the "running maximum" of Brownian motion, then $|B_t| \sim \max_{s\in[0,t]} B_s$.

4.5 Moment Generating Functions

So far, we have found two ways to extract information about an arbitrary distribution: the mean, and the variance. These are examples of *moments* of a distribution or random variable.

Definition 4.9

Given a random variable X, the **n**-th moment of X is $\mathbb{E}[X^n]$.

We can see that the mean of a distribution is simply its first moment. Moreover, the variance is related to the second moment by the formula $\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \text{Var}(X)$.

We now turn our attention to the problem of finding the moments of the normal distribution. In particular, consider $Z \sim \mathcal{N}(0, 1)$. Then when k is odd, we have

$$\mathbb{E}[Z^k] = \int_{-\infty}^{\infty} z^k \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz$$

But this is the product of an odd and even function, so it is odd and thus the integral is 0:

$$\mathbb{E}[Z^k] = 0$$

For even k, then the moment is not zero, and the integral can be evaluated with Feynman's trick. Instead, we will use *moment generating functions* to evaluate this.

Definition 4.10

Given a random variable X, then the **moment generating function** of X is $M_X(t) = \mathbb{E}[e^{tX}]$.

Example 4.3

If $X \sim \mathcal{N}(0, \sigma^2)$, then $M_X(t) = e^{t^2\sigma^2/2}$.

In particular, the moment generating function allows us to easily calculate moments of X:

$$M'_X(0) = \frac{d}{dt}|_{t=0}\mathbb{E}[e^{tX}] = \mathbb{E}[Xe^{tX}]|_{t=0} = \mathbb{E}[X]$$

More generally,

$$M_X^{(k)}(0) = \frac{d^k}{dt^k}_{t=0}\mathbb{E}[e^{tX}] = \mathbb{E}[X^k e^{tX}]|_{t=0} = \mathbb{E}[X^k]$$

So differentiating the moment generating function repeatedly allows us to calculate moment easily. Returning to our previous question, and allowing for other variances, suppose $X \sim \mathcal{N}(0, \sigma^2)$.

$$M_X(t) = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{tx - x^2/2\sigma^2}\, dx$$

Moment generating functions serve an important function in probability which is analogous to that of the Taylor series in analysis. For nice statistics f, we can express their expectations as some polynomial

$$\mathbb{E}[f(X)] = c_0 + c_1\mathbb{E}[X] + c_2\mathbb{E}[X^2] + \dots$$

In particular, as we generate more moments, we can better describe X, and in the limit, we have the following:

Theorem 4.11

The distribution of a random variable X is uniquely determined by its moments.

This means that if computations using the moment generating function line up with a function whose distribution we know, then the distributions must be the same.

Example 4.4

If $X \sim \mathcal{N}(0, \sigma^2)$ and $Y \sim \mathcal{N}(0, \eta^2)$, and $X \perp Y$, then the moment generating function of $X + Y$ is given by

$$M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}e^{tY}] = \underbrace{\mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}]}_{\text{by independence}}$$

From the previous example, we can fill this in:

$$M_{X+Y}(t) = e^{t^2\sigma^2/2}e^{t^2\eta^2/2} = e^{t^2(\sigma^2+\eta^2)/2}$$

But this is precisely the moment generating function of a variable with distribution $\mathcal{N}(0, \sigma^2 + \eta^2)$, so we must have $X + Y \sim \mathcal{N}(0, \sigma^2 + \eta^2)$.

Example 4.5

Let $X \sim \text{Exponential}(\lambda)$. The moment generating function of X is

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} f_X(x)dx = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx = \lambda \int_0^\infty e^{-(\lambda-t)x}dx$$
$$= \frac{\lambda}{\lambda - t}\int_0^\infty (\lambda - t)e^{-(\lambda-t)x}dx = \frac{\lambda}{\lambda - t}$$

This highlights the important note that a moment generating function may not exist everywhere, but in order to be practical, it must exist for an interval about 0 so that we can take the derivatives.

Recall from our discussion of normal distributions that if X, Y are normal and independent, then $X + Y$ is normal as well. the following is an example where this does not hold:

Example 4.6

Let $X \sim \mathcal{N}(0, 1)$ and η be ± 1, each with probability $1/2$. Suppose $X \perp \eta$. Then let $Y = \eta X$. Note that we have

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X \le y)\mathbb{P}(\eta = 1) + \mathbb{P}(X \ge -y)\mathbb{P}(\eta = -1)$$
$$= \frac{1}{2}\mathbb{P}(X \le y) + \frac{1}{2}\mathbb{P}(-X \le y)$$

Since X is symmetric, $X \sim -X$ and thus

$$F_Y(y) = \mathbb{P}(X \le y) = F_X(y)$$

So $F_X = F_Y$ and thus $X \sim Y$. Note also that X and Y are uncorrelated here:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - 0)(Y - 0)] = \mathbb{E}[XY] = \mathbb{E}[\eta X^2] = \underbrace{\mathbb{E}[\eta]\mathbb{E}[X^2]}_{X \perp \eta} = 0$$

However, remember that being uncorrelated does not imply independence. Importantly, we cannot conclude that $X + Y$ is normal, because it is not! Recall that a normal distribution is continuous, and thus the probability that it takes any given value must be 0. But we have

$$\mathbb{P}(X + Y = 0) = \mathbb{P}(X + \eta X = 0) = \mathbb{P}((1 + \eta)X = 0) = \mathbb{P}(\eta = -1) = \frac{1}{2}$$

Thus we could not have a fully continuous distribution as required by the normal distribution.

Remark

In fact, the distribution above is our first example of a mixed distribution. It assumes the value 0 with probability 1/2 (discretely), and a continuous normal distribution $\mathcal{N}(0, 4)$ with probability 1/2.

Recall that a collection of vectors $\overrightarrow{X} = \langle X_1, \ldots, X_d \rangle$ are *jointly Gaussian* if we have that, for all $\overrightarrow{\alpha} = \langle \alpha_1, \ldots, \alpha_d \rangle \in \mathbb{R}^d$, we have that $\overrightarrow{\alpha} \cdot \overrightarrow{X} = \alpha_1 X_1 + \ldots + \alpha_d X_d$ is Gaussian. In particular, we have that $\overrightarrow{\alpha} \cdot \overrightarrow{X} \sim \mathcal{N}(\overrightarrow{\alpha} \cdot \overrightarrow{\mu}, \overrightarrow{\alpha} \cdot \Sigma \overrightarrow{\alpha})$. Here, Σ is the covariance matrix, which has $\text{Cov}(X_i, X_j)$ as its ij-th entry.

Some notes about the covariance matrix is that it is symmetric, positive semidefinite, and has the variances $\text{Var}(X_i)$ along the diagonal. This matrix is a substitute in the multivariable case for the variance, such that a multivariate Gaussian distribution is parameterized by $(\overrightarrow{\mu}, \Sigma)$ rather than (μ, σ^2).

We can alternately express the covariance matrix as follows:

$$\Sigma = \begin{bmatrix} \text{Var}(X_1) & \ldots & \ldots & \text{Cov}(X_1, X_d) \\ \vdots & \text{Var}(X_2) & \ldots & \text{Cov}(X_2, X_d) \\ \text{Cov}(X_{d-1}, X_1) & \ldots & \ddots & \vdots \\ \text{Cov}(X_d, X_1) & \ldots & \ldots & \text{Var}(X_d, X_d) \end{bmatrix} = \mathbb{E}[(\overrightarrow{X} - \overrightarrow{\mu})(\overrightarrow{X} - \overrightarrow{\mu})^T$$

Let us now consider the moment generating function of a multivariate Gaussian variable $\overrightarrow{X}$. We have

$$M_{\overrightarrow{X}}(\overrightarrow{\alpha}) = \mathbb{E}[e^{\overrightarrow{X} \cdot \overrightarrow{\alpha}}] = e^{\overrightarrow{\alpha} \Sigma \overrightarrow{\alpha}/2} + \overrightarrow{\mu} \cdot \overrightarrow{\alpha}$$

Since we know that multivariate Gaussians are also jointly Gaussian, we also have

$$M_{\overrightarrow{\alpha} \cdot \overrightarrow{X}}(t) = \mathbb{E}[e^{t \overrightarrow{\alpha} \cdot \overrightarrow{X}}] = e^{t^2 \overrightarrow{\alpha} \cdot \Sigma \overrightarrow{\alpha}/2 + t(\overrightarrow{\mu} \cdot \overrightarrow{\alpha}}$$

4.6 Multivariate Brownian Motion

Now that we have understood how Gaussian distributions operate in multiple dimensions, we turn our attention to the study of Brownian motion in multiple dimensions. This is an important extension of our previous study, since Brownian motion occurs in physical phenomena in both 2 and 3 dimensions, and not simply walks along a line.

Example 4.7

Consider the random vector $\overrightarrow{X} = \langle B_s, B_t \rangle$ for some Brownian motion $(B_t)_{t \geq 0}$, and assume that $s < t$. Then for any $\overrightarrow{\alpha} \in \mathbb{R}^2$, we have

$$\overrightarrow{\alpha} \cdot \overrightarrow{X} = \alpha_1 B_s + \alpha_2 B_t = \alpha_1 B_s + \alpha_2 (B_t - B_s + B_s) = (\alpha_1 + \alpha_2) B_s + \alpha_2 (B_t - B_s)$$

But we know that $B_t - B_S \perp B_s$, so this is a linear combination of independent Gaussians. Thus, $\overrightarrow{\alpha} \cdot \overrightarrow{X}$ is also Gaussian. Since $\overrightarrow{\alpha}$ was arbitrary, we see that B_s, B_t are jointly Gaussian. Note that this idea holds if we extend further to n dimensions.

Example 4.8

Suppose we define $X_t \coloneqq B_t - tB_1$, with $0 < t < 1$. Then for any fixed t, X_t is a linear combination of two times in a Brownian motion, which we just showed are jointly Gaussian. Thus X_t is Gaussian. Then we can determine its distribution by calculating directly:

$$\mathbb{E}[X_t] = \mathbb{E}[B_t] - t\mathbb{E}[B_1] = \mathbb{E}[\mathcal{N}(0,t)] - t\mathbb{E}[\mathcal{N}(0,1)] = 0$$

We can also calculate the variance:

$$\mathrm{Var}(X_t) = \mathrm{Var}(B_t) + \mathrm{Var}(-tB_1) + 2\,\mathrm{Cov}(B_t, -tB_1) = t + t^2 + -2t(\min(t,1)) = t - t^2 = t(1-t)$$

So we see that $X_t \sim \mathcal{N}(0, t(1-t))$.

Chapter 5

Markov Chains

5.1 Elementary Markov Chains

Suppose we consider some general stochastic process $(X_n)_{n\geq 0}$ in discrete time which draws from a finite sample space $S = \{S_1, \ldots, S_N\}$. In order to fully describe this process, we must be able to calculate the joint distributions of any finite set of times in the stochastic process. In other words, given any $n \geq 0$ and any elements $x_0, \ldots, x_n \in \mathcal{S}$, we must be able to calculate $f_{X_0,\ldots,X_n}(x_0, \ldots, x_n)$. This is a very difficult calculation in general, but some of the techniques we have learned help us determine these values.

For instance, since we have a notion of these draws occurring in some temporal order, the law of multiplication helps us break this into small products:

$$\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) = \mathbb{P}(X_0 = x_0) \cdot \prod_{k=1}^{n} \mathbb{P}(X_k = x_k | X_0 = x_0 \ldots X_{k-1} = x_{k-1})$$

One issue with these products is that they have *high dimensionality*. To reduce the dimension of the problem, we use *Markov chains*. Markov chains, and more generally the Markov property, essentially says that the past and future are independent, and that the process will start over from each given state, independent of how it got there.

Definition 5.1

We say that $(X_n)_{n\geq 0}$ has the **Markov property** if for any k and any values $x_0, \ldots x_k \in \mathcal{S}$, we have

$$\mathbb{P}(X_k = x_k | X_0 = x_0, \ldots X_{k-2} = x_{k-2}, X_{k-1} = x_{k-1}) = \mathbb{P}(X_k = x_k | X_{k-1} = x_{k-1})$$

Here, $X_k = x_k$ represents the future, $X_{k-1} = x_{k-1}$ represents the present state, and all the X_i, $i \leq k-2$ terms are the past. So if $(X_n)_{n\geq 0}$ has the Markov property, then we can simplify our first equation to be

$$\mathbb{P}(X_0 = x_0) \cdot \prod_{k=1}^{n} \mathbb{P}(X_k = x_k | X_0 = x_0 \ldots X_{k-1} = x_{k-1}) = \prod_{k=0}^{n} \mathbb{P}(X_k = x_k | X_{k-1} = x_{k-1})$$

However, this doesn't fully capture the notion of "starting over." In order to do so, we want to require not only independence of different times, but also independence with respect to time. This is precisely the property that we have previously referred to as "stationarity."

Definition 5.2

We say $(X_n)_{n\geq 0}$ is **time homogeneous** or has stationarity if there exist $N^{@}$ numbers $p_{i,j}|i,j \in \mathcal{S}$ such that $\mathbb{P}(X_k = x_k|X_{k-1} = x_{k-1}) = p_{x_{k-1},x_k}$ for all k.

The values $p_{i,j}$ here represent the probability at any given time of moving from state i to state j. Thus we have reduced the previous formula to

$$\prod_{k=0}^{n} \mathbb{P}(X_k = x_k|X_{k-1} = x_{k-1}) = \mathbb{P}(X_0 = x_0)p_{x_0,x_1}p_{x_1,x_2}\cdots p_{x_{k-1},x_k}$$

This means that rather than having to know every possible conditional probability for any combination of values, we only need to know the probability that X_0 assumes each of the possible initial values, and the transition probabilities $p_{x,y}$. For this finite case, that means we only need to calculate $N^2 + N$ values, rather than an infinite number of values.

Definition 5.3

A stochastic process in discrete time with finite state space $S = \{s_1, \ldots, s_n\}$ that has the Markov property and is time homogeneous is a **time homogeneous finite Markov chain**.

From the discussion above, we only need two pieces of information to identify a Markov chain. First, we need to know the initial distribution $\mu = (\mu_{s_1}, \ldots, \mu_{s_n})$, which is a $1 \times n$ row vector such that μ_{s_i} is just the probability that the starting value is s_i, such that $\mu_{s_i} = \mathbb{P}(X_0 = s_i)$. In order for these probabilities to make sense, we would mandate that $\mu_i \geq 0$ for all i and that $\sum \mu_i = 1$.

We will also record all the transition probabilites in a **stochastic matrix** $P = (p_{x_i,x_j})_{x_i,x_j\in\mathcal{S}}$, where $p_{x_i,x_j} = \mathbb{P}(X_k = x_j|X_{k-1} = x_i)$ represents the probability of moving from the row state x_i to the column state x_j. In order for this to make sense, we must have $p_{x_i,x_j} \geq 0$ for all x_i, x_j and $\sum_j p_{ij} = 1$ when summing over a row (but not necessarily a column). We conventionally refer to P itself as the Markov chain, since μ only determines the initial state, and P governs all the changes throughout time.

Before we move on, let us make some notes about notation. Given some initial distribution μ, we write $\mathbb{P}(X_n = x_n)$, or more generally $\mathbb{P}(X_n \in A)$ for some $A \subseteq S$, to be $\mathbb{P}_\mu(X_n = x_n)$ and $\mathbb{P}_\mu(X_n \in A)$ to explicitly specify the original state of the Markov chain. Moreover, when $\mu = (0, \ldots 0, 1, 0 \ldots, 0)$, such that it is guaranteed that we start in state s_x, then we write $\mathbb{P}_{s_k} = \mathbb{P}_\mu$ to specify the known starting value, such that $\mathbb{P}_{s_k}(X_n = x_j) = \mathbb{P}(X_n = x_j|X_0 = s_k)$. Lastly, we will write the transition properties in the shorthand p_{ij}. Moreover, when there is understood to be some fixed ordering of $S = \{s_1, \ldots, s_n\}$, we will sometimes refer to the state s_i as simply i.

Suppose we want to calculate the general distribution of X_1, given the initial distribution μ. Then this is

$$\mathbb{P}_\mu(X_1 = j) = \sum_{k\in\mathcal{S}} \underbrace{\mathbb{P}_\mu(X_0 = k)}_{\mu_k} \underbrace{\mathbb{P}_\mu(X_1 = j|X_0 = k)}_{p_{kj}}$$

In order to calculate this, consider what happens when we calculate

$$\mu P = \begin{bmatrix}\mu_1 & \dots & \mu_n\end{bmatrix} \begin{bmatrix} p_{11} & \dots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \dots & p_{nn}\end{bmatrix} = \begin{bmatrix}\sum_j \mu_j p_{j1} & \dots & \sum_j \mu_j p_{jn}\end{bmatrix}$$

So we see that the probability $\mathbb{P}_\mu(X_1 = j)$ is precisely given by the jth element of μP. So the row vector μP essentially gives us the "initial distribution" for time 1 rather than time 0.

If we then consider X_2, we have the following

$$\begin{aligned}\mathbb{P}_\mu(X_2 = j) &= \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{S}} \mathbb{P}_\mu(X_0 = i)\mathbb{P}_\mu(X_1 = k|X_0 = i) \underbrace{\mathbb{P}_\mu(X_2 = j|X_0 = i, X_1 = k)}_{=\mathbb{P}(_\mu(X_2=j|X_1=k)} \\ &= \sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{S}} \mu_i p_{ik} p_{kj} = \sum_{k\in\mathcal{S}} (\mu P)_k p_{kj}\end{aligned}$$

By the previous calculation, this value is the jth element of $\mu PP = \mu P^2$. Then by induction, we have the following:

Theorem 5.1

Let $(X_n)_{n\geq 0}$ be a Markov chain with initial distribution μ and stochastic matrix P. Then for any value j and any time n, we have

$$\mathbb{P}_\mu(X_n = j) = (\mu P^n)_j$$

where $(\mu P^n)_j$ is the jth element of μP^n.

Example 5.1

Suppose a frog hops between two lily pads. Before each hop, the frog flips a p-coin if the frog is on pad 1, and a q-coin if the frog is on pad 2 (with not $p+q$ necessarily 1). If the frog flips heads, it switches pads, otherwise, it stays on the same pad.

Given this information, we can write the stochastic matrix as follows:

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22}\end{bmatrix} = \begin{bmatrix} 1-p & p \\ q & 1-q\end{bmatrix}$$

Moreover, the distribution of the initial pad the frog is on must take the form

$$\mu = \begin{bmatrix} r & 1-r\end{bmatrix}$$

For some $r \in [0,1]$. Then the distribution of the pad the frog is on at time n is

$$\mathbb{P}(X_n = 1) = (\mu P^n)_1, \mathbb{P}(X_n = 2) = (\mu P^n)_2$$

There is in general no guarantee that any single path through states in a Markov chain converges to some single state. In the example above, if $p, q > 0$, then the frog's path almost never converge to one lily pad. However, it is indeed possible for the distribution at time n, given by $\mu^{(n)}$, to converge.

Suppose that $\mu^{(n)}$ converges to some row vector π. Then we must have

$$\pi = \lim_{n\to\infty} \mu^{(n)} = \lim_{n\to\infty} \mu P^{n+1} = P \lim_{n\to\infty} \mu P^n = P\pi$$

So π can only be a limit for the distribution if it satisfies $\pi = \pi P$. We can think of these distributions as fixed points or steady states within the space of distributions.

Definition 5.4

A $1 \times n$ row vector π which has $\pi = P\pi$ and has $\pi_i > 0$ for all i and $\sum_i \pi_i = 1$ is a **stationary distribution** for P.

Note that we do not have guarantees of uniqueness here - different initial states may accumulate around different steady states. Moreover, we have not yet shown that every chain converges.

Example 5.2

Consider the 2×2 case. Suppose that $\pi = \begin{bmatrix}\pi_1 & \pi_2\end{bmatrix}$. Also, the stochastic matrix must take the form $\begin{bmatrix}1-p & p \\ q & 1-q\end{bmatrix}$. So if π is a stationary distribution, it must be the case that

$$\pi = \begin{bmatrix}\pi_1 & \pi_2\end{bmatrix} = \begin{bmatrix}1-p & p \\ q & 1-q\end{bmatrix} \begin{bmatrix}\pi_1 & \pi_2\end{bmatrix} = \begin{bmatrix}\pi_1(1-p) + \pi_2 q \\ \pi_1 q + \pi_2(1-q)\end{bmatrix}$$

Which implies that

$$0 = -p\pi_1 + q\pi_2 = -p\pi_1 + q(1-\pi_1) \implies \pi_1 = \frac{q}{p+q}$$

and similarly

$$\pi_2 = \frac{p}{p+q}$$

So

$$\pi = \begin{bmatrix}\frac{q}{p+q} & \frac{p}{p+q}\end{bmatrix}$$

if the limit exists.

We can interpret this result by saying that the kth entry of π is the fraction of time, on average, that the chain spends in state k.

Let us now handle the question of convergence in the 2×2 case. Although we do not know whether there is convergence, we do know that if there is convergence, we must have π as the limit. Thus, define $\Delta^{(n)} := \mu^{(n)} - \pi$. So we want to show that $\Delta^{(n)} \to 0$.

For any n, we can use the properties of the Markov chain to replace matrix multiplication. We have

$$\Delta_1^{(n+1)} = \mu_1^{(n+1)} + \pi_1 = \mu_1^{(n)}(1-p) + \mu_2^{(n)} q - \frac{q}{p+q}$$

using the fact that $\mu_2^{(n)} = 1 - \mu_1^{(n)}$,

$$\begin{aligned}\Delta_1^{(n+1)} &= \mu_1^{(n)}(1-p-q) + q\left[1 - \frac{1}{p+q}\right] = \mu_1^{(n)}(1-p-q) - q\left[\frac{1-p-q}{p+q}\right] \\ &= (1-(p+q))(\mu_1^{(n)} - \underbrace{\frac{q}{p+q}}_{\pi_1}) = (1-p-q)\Delta_1^{(n)}\end{aligned}$$

Thus we must have $\Delta^{(n)} = \mu^{(n)} - \pi$. Moreover, from this calculation,

$$\Delta^{n+1} = (1-p-q)\Delta^{(n)} = (1-(p+q))^n \Delta^{(0)} = (1-(p+q))^n(\mu - \pi)$$

Then $\mu^{(n)}$ converges to π if $|1-(p+q)| < 1$.

Let us also consider some corner cases. If $p = q = 0$, then the previous formula for π is invalid by division by 0. Moreover, we can see intuitively that the initial distribution will never change, so that every distribution is a stationary distribution. This makes sense, because our matrix in this case is given by

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Additionally, if $p = q = 0$, then the chain will never converge to our stationary point π, unless $\mu = \pi$.

Example 5.3

Suppose we extend the previous lily pad example to any larger number of pads. However, we mandate that there are two groups of lily pads, and that there probability of moving between groups is 0. We impose no constraints on the probability of moving within a group. Then the matrix is given by

$$P = \begin{bmatrix} \tilde{P} & O \\ O & \tilde{Q} \end{bmatrix}$$

where $\tilde{P}$ and $\tilde{Q}$ are the stochastic matrices of the restrictions to each group.

We call the ability to break the chain into two smaller chains a "reduction" of the chains. When a chain can be reduced, we see that convergence to the steady state π fails unless $\mu = \pi$.

Definition 5.5

Let $i, j \in \mathcal{S}$ be two states. Let $M > 0$ be an integer. Then we denote by $p_{ij}^{(M)}$ the probability $\mathbb{P}(X_M = j | X_0 = i)$. Moreover, $p_{ij}^{(M)}$ is the i, jth entry of P^M.

Intuitively, this is the probability that we reach state j from state i in exactly M steps. Then to be irreducible, we want there to be no isolated subgroups. That is, we should eventually be able to get between any two states.

Definition 5.6

A chain P is called **irreducible** if for any states $i, j \in \mathcal{S}$, there exists some $M > 0$ such that $p_{ij}^{(M)} > 0$.

Note that M may depend on i, j in general.

In another edge case, when $p = q = 1$, then we have "periodic" behavior cycling through the states. For instance, if we alternate between two states, then a distribution which begins entirely in one state will never converge to any stationary distribution.

Intuitively, periodicity occurs when there are states such that travel between them over exactly M steps depends on M. For instance, in the alternating case, we would only be able to travel between i, j in an odd number of steps. This restriction is what we call periodicity.

Definition 5.7

Let $i \in \mathcal{S}$ be a state. Consider the set $\mathcal{R}_i \coloneqq \{n \geq 1 | p_{ii}^{(n)} > 0\}$ of all the numbers such that there is a path between i and itself. Then the **period** of i is $d_i \coloneqq \gcd(\mathcal{R}_i)$.

Definition 5.8

A chain P has period $d \in \mathcal{N}$ if every state $i \in \mathcal{S}$ has period $d_i = d$.

Definition 5.9

A chain P is **aperiodic** if it has period 1.

For instance, if there is always a nonzero probability of remaining in any state, then the chain is aperiodic.

Example 5.4

Consider the set $\mathbb{Z}_4$, and consider a random walk on this graph (visualized as a walk around a circle with four points). Is this walk periodic?

Take $i = 1$, and consider $\mathcal{R}_1$. Note that after every step, the parity of the state must change. So it takes an even number of steps to return to 1: $\mathcal{R}_1 = \{2, 4, 6, \ldots\}$. Thus $d_1 = 2$. Similarly, $d_i = 2$ for every i here. So the period of the walk is 2 and the walk is periodic.

However, if we consider $\mathbb{Z}_3$ instead, we get $\mathcal{R}_1 = \{2, 3, 4, \ldots\} = \mathbb{N} \setminus \{1\}$. So $d_1 = 1$ and this walk is aperiodic.

Note that our definition above only defines the period when every state has the same period. Thus, some states are not aperiodic, but also do not have a period. This ends up being sufficient for our purposes, since we only consider aperiodic chains which are also irreducible.

Lemma

Let P be an irreducible chain. Then every state has the same period, so there is some d such that $d_i = d$ for every $i \in \mathcal{S}$.

Then every time we discuss an irreducible chain, we can assume it has a well-defined period.

Lemma

If P is finite, irreducible, and aperiodic, then there is some $M > 0$ such that $p_{ij}^{(m)} > 0$ for any $m > M$.

The result of this lemma essentially says that although some states may initially be inaccessible from others, the graph of reachable states will eventually be fully connected.

Example 5.5

For any odd integer $n = 2k + 1$, if P is a random walk on the circular graph of $\mathbb{Z}_n$, then there is a way to pass between any two states in an odd number of steps or an even number of steps. Since we can always increase this number of steps by 2 by taking two steps in opposite directions, there is a path of any sufficiently large length between any two states. Precisely, this number is $M = k$.

Definition 5.10

Let $i \in \mathcal{S}$ be a state and $(X_n)_{n \geq 0}$ be a chain. Then define $T_i = \min\{k \geq 0 | X_k = i\}$ to be the first **hitting time** of i. Define $T_i^+ := \min\{k \geq 1 | X_k = i\}$ to be the first **return time** of i.

Definition 5.11

A state i is **recurrent** if $\mathbb{P}_i(T_i^+ < \infty) = 1$. That is, if the chain starts in state i, then it eventually returns with probability 1. If this is not the case, then i is **transient**.

Proposition 5.2

If P is a finite irreducible chain beginning in state i, $\mathbb{E}_i[T_i^+] < \infty$ and there exists a unique stationary distribution π with the ith entry given by $\pi_i = \frac{1}{\mathbb{E}_k[T_k]}$.

Proof. Fix some state $i \in \mathcal{S}$. Let $S_i^{+,1}, S - i^{+,2}, \ldots$ be the inter-return times to the state i, given that we started in state i. Note that $S_i^{+,1} = T_i$. Moreover, by the Markov property, these are independent and identically distributed. Finally, consider $N_m \coloneqq S_i^{+,1} + \ldots + S_i^{+,m}$ to be the time it takes to return m times. Then the total count of visits is m. Then since π_i is the fraction of time spent in state i in the long run, we have

$$\pi_i = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1_{\{X_k = i\}} = \lim_{m\to\infty} \frac{m}{N_m} = \frac{1}{\mathbb{E}_i[T_i^+]}$$

where the second equality follows from the irereducibility argument, which gives us recurrence. Thus the theorem is proved. □

Example 5.6: Knight's Walk

Consider a knight traveling around a chessboard. How long, on average, does it take to return to the starting square?

We first order the squares on the board and create an unnormalized stationary distribution $\tilde{\pi}$. Recalling that the entries represent the relative amount of time spent at a given square, and that the ability to get to a square depends on the number of squares which are connected to it by a knight's move (the degree of the square within the graph of knight's moves). So our unnormalized vector has $\tilde{\pi}_i = \deg(i)$. To normalize, we then divide by twice the number of edges, to get $\pi_i = \frac{\deg(i)}{2|E|}$. Then we can use the above formula to get

$$\mathbb{E}_i[T_i] = \frac{1}{\pi_i} = \frac{2|E|}{\deg(i)} = \frac{336}{\deg(i)}$$

For instance, if the knight starts in a corner, then the degree is 2, and thus $\mathbb{E}_{\text{corner}}[T_{\text{corner}}^+] = 168$.

Theorem 5.3

If a finite chain P is irreducible and aperiodic, then there exists some stationary distribution π such that for all initial distributions μ, $\mu^{(n)} = \mu P^n \to \pi$.

Note that the irreducible and aperiodic condition is necessary but not sufficient. For instance, consider the chain

$$P = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

Then every distribution converges to the distribution $\pi = \begin{bmatrix} 1 & 0 \end{bmatrix}$, but the chain is reducible as there is no way to reach state 2 from state 1 (i.e. the right column is 0 for any power P^k).

This discussion allows us to approach the law of large numbers when we don't have an independence assumption.

Definition 5.12

We call a chain **ergodic** if there exists π stationary such that for any function $F : \mathcal{S} \to \mathbb{R}$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k) = \sum_{i\in\mathcal{S}} F(i)\pi_i$$

Theorem 5.4: Ergodic Theorem

If a finite chain P is irreducible and aperiodic, then it is ergodic.

In other words, the average of a function F over time (n considered as time), is the same as as the average of the function over space ($\mathcal{S}$ considered as space). This idea is very powerful and used in physics and Fourier analysis.

Remark

The above three theorems hold if $\mathcal{S}$ is finite. They also hold for countably infinite $\mathcal{S}$, so long as $\mathbb{E}_i[T_i] < \infty$.

Recall that we can think of π_i as the long-run fraction of time spent in state i. In other words, we have

$$\pi_i = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1_{\{X_n=i\}}$$

We can also consider the inter-return times, that is, $S_i^{+,1} = T_i, S_i^{+,2}, \ldots$, and we write $N_m = \sum_{k=1}^{m} S_i^{+,k}$. When we have $\mathbb{E}_i[T_i^+] < \infty$, then m goes to infinity as n does, so we can rewrite as

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1_{\{X_n=i\}} = \lim_{m\to\infty} \frac{m}{N_m} = \lim_{m\to\infty} \frac{1}{\frac{S_i^{+,1}+\ldots+S_i^{+,m}}{m}}$$

By the Markov property, each of these inter return times is independent, and all of the terms $S_i^{k,+}$ with $k > 1$ are distribution according to T_i^+, so using the law of large numbers we get

$$\lim_{m\to\infty} \frac{1}{\frac{S_i^{+,1}+\ldots+S_i^{+,m}}{m}} = \frac{1}{\mathbb{E}_i[T_i^+]}$$

(Note that the first term is distributed according to T_i instead of T_i^+, since we may not start in state i; however, this distinction does not matter in the limit).

5.2 First Step Analysis

Suppose we generalize the notion of hitting times for a state to hitting times for a set.

Definition 5.13

Let $A \subseteq \mathcal{S}$. Then define the first hitting time of A to be

$$T_A(X_0, X_1 \ldots) = T_A \coloneqq \min\{n \geq 0 | X_n \in A\}$$

Of course, when $x \in A$, then we start in A so $\mathbb{P}_X(T_A = 0) = 1$. On the other hand, if $x \notin A$, then we are guaranteed to have to take at least one step. So we can rewrite this as

$$T_A(X_0, X_1, \ldots) = 1 + T_A(X_1, X_2, \ldots)$$

Let us define some convenience functions:

$$\begin{aligned} p(x) &\coloneqq \mathbb{P}_x(T_A < \infty) \\ g(x) &\coloneqq \mathbb{P}_x(X_{T_A} = b) \\ m(x) &\coloneqq \mathbb{E}_x[\sum_{k=0}^{T_A - 1} g(X_k)] \end{aligned}$$

Then when we start outside of A, we can write

$$m(x) = g(x) + \mathbb{E}_x[\sum_{k=1}^{T_A - 1} g(X_k)$$

Conditioning on the first step we take, this is

$$m(x) = g(x) = \sum_{y \in \mathcal{S}} \mathbb{P}_x(X_1 = y) \mathbb{E}_x \left[\sum_{k=1}^{T_A - 1} g(X_k) | X_1 = y \right] = g(x) + \sum_{y \in \mathcal{S}} p_{xy} m(y)$$

with the boundary condition

$$m(x) = 0, x \in A$$

5.3 Classification of States

We will now investigate further ways that we can classify different states in a Markov chain.

Definition 5.14

Given a chain P, we say that two states $i, j \in \mathcal{S}$ **communicate**, denoted $i \leftrightarrow j$, if there exists $m, n \geq 0$ such that $p_{ij}^{(m)}, p_{ji}^{(n)} > 0$.

Intuitively, this means you can pass from one state to the other and back. We note that this is an equivalence relation on $\mathcal{S}$. Thus, we can define the equivalence classes:

Definition 5.15

Given some state $i \in \mathcal{S}$, the **communication class** of i is defined as $[i] = \{j \in \mathcal{S} : i \leftrightarrow j\}$.

Recall that a state $i \in \mathcal{S}$ is recurrent if $\mathbb{P}_i(T_i^+ < \infty) = 1$, and is transient otherwise. We can apply siilar language to communication casses:

Definition 5.16

A communication class $\mathcal{R}$ is **recurrent** if $p_{ij} = 0$ for all $i \in \mathcal{R}, j \notin \mathcal{R}$.

In other words, there is no way to leave the communication class.

Definition 5.17

A communication class $\mathcal{T}$ is **transient** if it is not recurrent; that is, there exist $i \in \mathcal{T}, j \notin \mathcal{T}$ such that $p_{ij} > 0$.

Note that this definition implies you can not return to the communication class, since otherwise that would mean that j would be in the transient class. So you can leave a transient class, but not return to it.

Proposition 5.5

If $i \in \mathcal{R}$ for some recurrent communication class, then $\mathbb{P}_i(T_i^+ < \infty) = 1$ (all states are recurrent) and for any $j \in \mathcal{R}$, $\mathbb{P}_i(X_n = j \text{ for infinitely many } n) = 1$.

Proposition 5.6

If $\mathcal{T}$ is transient, any Markov chain starting in $\mathcal{T}$ will eventually leave it and never return.

Definition 5.18

A chain P is irreducible if it has exactly one communication class $\mathcal{S} = \mathcal{R}$ (which must be recurrent).

Example 5.7: Gambler's Ruin

Consider a biased random walk with a p-coin, such that we stop the walk when it reaches the lower bound of a or the upper bound of b. This stopping condition means that any chain is "absorbed" when it hits the boundary $\mathcal{A} = \{a, b\}$. So the recurrent

communication classes are $\mathcal{R}_1 = \{a\}$ and $\mathcal{R}_2 = \{b\}$ and the transient class is the remainder: $\mathcal{T} = \{a+1, \ldots, b-1\}$.

Example 5.8

Suppose we have a six state space $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ and a chain with transition matrix

$$P = \begin{bmatrix} 1/3 & 2/3 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/3 & 1/6 & 0 \\ 0 & 0 & 3/5 & 1/5 & 1/5 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 2/3 & 1/3 \end{bmatrix}$$

Note that the diagonal entries are nonzero, so every state can return to itself after an iteration of the chain. Let us consider the states which can be passed to:

$$\begin{cases} 1 \to \{1, 2\} \\ 2 \to \{1, 2\} \\ 3 \to \{1, 2, 3, 4, 5\} \\ 4 \to \{3, 4, 5\} \\ 5 \to \{5, 6\} \\ 6 \to \{5, 6\} \end{cases}$$

If we represent this as a graph, we can draw paths through the graph to determine the recurrent and transient states and classes. Note that the classes $\{1, 2\}$ and $\{5, 6\}$ are both communication classes, since they have arrows in both directions within the class, but there are no arrows out of the class. So the recurrent classes are $[1] = [2], [5] = [6]$. This directly implies that 1,2,5,6 are all recurrent. Lastly, states 3 and 4 communicate with each other, so $\{3, 4\}$ is a communication class, and we can clearly leave this class, so it is transient. Then 3,4 are both transient states.

Theorem 5.7

For any chain P with finite state space, we can partition $\mathcal{S}$ into a disjoint union of recurrent communication classes $\mathcal{R}_1, \ldots, \mathcal{R}_m$ and transient communication classes $\mathcal{T}_1, \ldots, \mathcal{T}_n$.

If we do this and then reorder our states so that the recurrent classes are together and come before the transient classes, then our matrix will take the form of a block matrix:

$$P = \begin{bmatrix} P_1 & O & \ldots & O \\ O & P_2 & \ldots & O \\ \vdots & \vdots & \ddots & \vdots \\ * & * & * & * \end{bmatrix}$$

where the stars represent the rows for the transient classes (since they can enter other classes, but cannot be entered into by recurrent classes). We call this a **substochastic matrix**.

By partitioning our general chain in this way, we can apply the theorems we found about specific chains, restricted to each recurrent class. That is, if we think about the submatrices $P_1, P_2, \ldots$ as subchains, then they will behave like an irreducible Markov chain within their respective recurrent classes.

Note that if we know some state $i \in \mathcal{S}$ is transient, then attempting to compute the expected return time to $\mathbb{E}_i[T_i^+]$ will always give ∞. Thus, we need to ignore the transient states. On the other hand, if i is recurrent, then computing $\mathbb{E}_i[T_i]$ allows us to determine the stationary distribution.

Continuing the above example,

Example 5.9

We want to compute $\mathbb{P}_3(T_{\{12\}} < \infty)$. In other words, given that we start in the transient state 3, what is the probability we end up in the class [1]?

using first step analysis, define $r(x) = \mathbb{P}_x(T_{\{1,2\}} < \infty)$. Clearly, $r(1) = r(2) = 1$ and $r(5) = r(6) = 0$. By looking at the matrix, $r(3) = 1/6r(1) + 1/6r(2) + 1/6r(3) + 1/3r(4) + 1/6r(5)$. Similarly, we compute $r(4)$ and solve to get $r(3) = 4/7$.

5.4 Countable Markov Chains

We will now extend our discussion of Markov chains so far to allow for countable state spaces $\mathcal{S}$.

Example 5.10

Consider a biased random walk reflected at 0, such that $\mathcal{S} = \mathbb{N}$. Then if we consider any state besides 0, it will transition into other states as follows:

$$p_{ij} = \begin{cases} p, j = i + 1 \\ 1 - p, j = i - 1 \\ 0 \text{ otherwise} \end{cases}$$

In the case $i = 0$, any flip will be "reflected" into 1, so

$$p_{0j} = \begin{cases} 1, j = 1 \\ 0 \text{ otherwise} \end{cases}$$

We will now recast some of the properties we investigated in the finite case in a manner that will allow us to convert into countable chains.

Theorem 5.8

Let P be an irreducible chain. Then there exists a recurrent states $i \in \mathcal{S}$ if and only if all states are recurrent and for any states $i, j \in \mathcal{S}$, the expected number of visits to j starting from i is infinite:

$$\mathbb{E}_i\left[\sum_{k=0}^{\infty} 1_{\{X_k=j\}}\right] = \infty$$

Theorem 5.9

Let P be irreducible. Then P is transient if and only if for any $i, j \in \mathcal{S}$,

$$\mathbb{E}_i\left[\sum_{k=0}^{\infty} 1_{\{X_k=j\}}\right] < \infty$$

We should note that although we have defined both using statements about all pairs of states, irreducibility shows that these two cases are the only two cases (and are disjoint).

Note that we can use linearity and expectation of indicators in the equation above to find that this is

$$\sum_{k=0}^{\infty} \mathbb{P}_i(X_k = j) = \sum_{k=0}^{\infty} p_{ij}^{(k)}$$

Moreover, $p_{ij}^{(k)} \to \pi_j$ in the limit.

We will now sketch a proof for the above two theorems.

Proof. Fix a state $i \in \mathcal{S}$ and let V_i be the number of visits to state i:

$$V_i = \sum_{k=0}^{\infty} 1_{\{X_k=i\}}$$

Define q to be the probability that i is eventually returned to:

$$q \coloneqq \mathbb{P}_i(T_i^+ < \infty)$$

Then letting $p = 1 - q$, this is the probability that the chain never returns to state i. I claim that $V_i \sim \text{Geom}(p)$.

To see this, note that the event $V_i = m$ is equivalent to the chain returning to i $m-1$ times, and subsequently never returning. This means for the first $m-2$ returns, we have a return afterward, and the probability of this is given by q. Then the probability of never returning after is p:

$$\mathbb{P}_i(V_i = m) = q^{m-2}p$$

Then we have

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = \mathbb{E}_i[V_i] = \frac{1}{p} = \frac{1}{1-q}$$

So this expectation is finite whenever $q < 1$ (there is a chance of leaving), then the state is transient. This makes sense, since as time goes to infinity, it becomes increasingly unlikely that we hit p at least once. If $q = 1$ (we are guaranteed to stay), then the expectation is infinite and the state is recurrent. □

To summarize, we find that a chain is transient when the expected number of revisits is finite, and recurrent when the expectation is infinite. This extends neatly to any two states, since we have the irreducibility assumption.

This means that when P is transient, we have $\mathbb{E}_i[T_i^+] = \infty$, so $\pi_i = \frac{1}{\mathbb{E}_i[T_i^+]} = 0$. So the interpretation is that the chain spends none of the long-run fraction of time in a transient state. Note, however, that this can also happen for recurrent chains.

Definition 5.19

An irreducible chain P is **null recurrent** if, given any state i, the expected number of visits is infinite but the long-run time spent in i is zero:

$$\begin{cases} \sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty \\ \lim_{n\to\infty} p_{ii}^{(n)} = 0 \end{cases}$$

P is **positive recurrent** if, given any i, the expected number of visits is infinite but the long-run time spent in i is positive:

$$\begin{cases} \sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty \\ \pi_i = \lim_{n\to\infty} p_{ii}^{(n)} > 0 \end{cases}$$

The key fact is that for an infinite state space $\mathcal{S}$, positive recurrent chains enjoy all the properties of finite chains.

Theorem 5.10: Kac's Theorem

Let P be a chain on an infinite state space $\mathcal{S}$.

1. An irreducible chain P is positive recurrent if and only if $\mathbb{E}_i[T_i^+] < \infty$ for all i and there exists a unique stationary distribution such that $\pi_i = \frac{1}{\mathbb{E}_i[T_i^+]}$.
2. If P is irreducible, aperiodic, and positive recurrent, there exists π such that given any initial distribution μ,
$$\mathbb{P}_\mu(X_n = i) = (\mu P^n) \to \pi$$
3. For any $F : \mathcal{S} \to \mathbb{R}$,
$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k) = \sum_{i\in\mathcal{S}} F(i)\pi_i = \mathbb{E}_\pi[F(X_0)]$$

Example 5.11

Returning to the biased random walk, then if we start in some state $x \geq 1$ and ask the probability that we reach some fixed N before 0 (gambler's ruin) is

$$\mathbb{P}_x(X_{T_{0,N}} = N)$$

When $p = \frac{1}{2}$, this is

$$\frac{x-0}{N-0} = \frac{x}{N}$$

When $p \neq \frac{1}{2}$, this is

$$\frac{(\frac{q}{p})^x - 1}{(\frac{q}{p})^N - 1}$$

Given this, the probability that we never hit x is the limit as N goes to infinite:

$$\mathbb{P}_x(T_0 = \infty) = \begin{cases} \lim \frac{x}{N} = 0, p = 1/2 \\ \lim \frac{(\frac{q}{p})^x - 1}{(\frac{q}{p})^N - 1} = 1 - (\frac{q}{p})^x, p > 1/2 \\ \lim \frac{(\frac{q}{p})^x - 1}{(\frac{q}{p})^N - 1} = 0, p < 1/2 \end{cases}$$

Then, starting at x, the probability that we *do* hit 0 is

$$\mathbb{P}_x(T_0 < \infty) = 1 - \mathbb{P}_x(T_0 = \infty) = \begin{cases} 1, p \leq 1/2 \\ \frac{q}{p}^x, p > 1/2 \end{cases}$$

So we can only hope to have a stationary distribution when $p \leq 1/2$, since otherwise we have transience. However, for $p = 1/2$, we can show that there is no stationary distribution, since $\pi_i = 0$, so we have null recurrence. Lastly, we have positive recurrence when $p < 1/2$, and in this case the distribution is

$$\pi_x = c(\frac{q}{p})^x = \frac{p}{p-q}(\frac{q}{p})^x$$

5.5 Branching Processes

Suppose we have some population of individuals, and after some fixed time period, each individual dies and is replaced with a number of individuals according to some random variable N. Suppose that each of these replacements is independent as well. Then after another period of time, or round, this occurs again for each of the new individuals. This is an example of what we call a branching process.

Definition 5.20

The following is the definition of a **branching process** known as a *Galton-Watson tree*. Suppose a random variable N takes nonnegative integer values. Suppose a certain population has X_k individuals at round k. Let $N_i^{(k)}$ denote the number of individuals which replace individual i, such that $N_i^{(k)} \sim N$ for all i, k. Then we recursively define $X_{k+1} = N_1^{(k)} + \ldots + N_{X_k}^{(k)}$ to be the number of individuals in round X_{k+1}. If we fix the number of starting individuals X_0, then this defines a Galton-Watson tree. In particular, we adopt the convention $X_0 = 1$ unless stated otherwise.

Note that we can interpret this as a Markov chain with countable state space $(0, 1 \ldots,)$. Note also that 0 is an absorbing state. Moreover, for any $k > 0$, because each individual branches independently of the others, the probability that there are k individuals in one round and 0 in the next is given by $\mathbb{P}(N = 0)^k$.

Definition 5.21

Suppose a random variable X takes nonnegative integer values. Then the **generating function** of X is

$$\phi_X(t) \coloneqq \mathbb{E}[t^X]$$

This can also be represented as the formal power series

$$\sum_{k=0}^{\infty} t^k \mathbb{P}(X = k)$$

Proposition 5.11

Consider a Galton-Watson tree with branching variable N. Then the *extinction probability* $\eta \coloneqq \lim_{n\to\infty} \mathbb{P}(X_n = 0)$ is a fixed point of $\phi_N(x)$.

Proof. Suppose that the individual branches into k individuals. Then each new individual defines a new tree, which goes extinct independent of the others. So $\mathbb{P}(\text{extinction}|X_1 = k) = \eta^k$. Then define η_i as the probability of extinction by round i. Thus for any n, we have

$$\eta_{n+1} = \sum_{k=0}^{\infty} \mathbb{P}(X_{n+1} = 0|X_1 = k)\mathbb{P}(X_1 = k)$$

But by the independence argument we have

$$\mathbb{P}(X_{n+1} = 0|X_1 = k) = \eta_n^k$$

So

$$\eta_{n+1} = \sum_{k=1}^{\infty} \eta_n^k \mathbb{P}(N = k) = \phi_N(\eta_n)$$

Since $\eta = \lim_{n\to\infty} \eta_n$, and ϕ_N is continuous, we have

$$\eta = \lim_{n\to\infty} \eta_{n+1} = \lim_{n\to\infty} \phi_N(\eta_n) = \phi_N(\lim_{n\to\infty} \eta_n) = \phi_N(\eta)$$

So η is a fixed point of ϕ_N. □

Theorem 5.12

The extinction probability of a Galton-Watson tree with branching variable N is the smallest fixed point of ϕ_N in the interval $[0, 1]$.

Proof. Let η^* be the smallest fixed point of ϕ_N in $[0, 1]$. We have

$$\eta_0 = 0 \leq \eta^*$$

From the proof of the previous proposition, we have

$$\eta_1 = \phi_N(\eta_0) = \phi_N(0)$$

Since $\mathbb{P}(N = k) \geq 0$ for all k, it can be seen that ϕ_N is increasing on $[0, \infty)$. So

$$\phi_N(0) \leq \phi_N(\eta^*) = \eta^*$$

By induction this holds for further η_n, so

$$\eta = \lim_{n\to\infty} \eta_n \leq \eta^*$$

but η is a fixed point in $[0, 1]$ so it must be at least η^*. Thus $\eta = \eta^*$. □

By the arguments made previously, we can adjust for any number of starting individuals by raising the power of η, so that the new extinction probability is $\eta' = \eta^{X_0}$.

Proposition 5.13

For a Galton-Watson tree $(X_n)_{n\geq 0}$ with branching variable N, $\mathbb{E}[X_n] = \mathbb{E}[N]^n$.

Proof. Induct. We have $\mathbb{E}[X_1|X_0 = 1] = \mathbb{E}[N]$. If $\mathbb{E}[X_n] = \mathbb{E}[N]^n$, then by linearity,

$$\mathbb{E}[X_{n+1}] = \mathbb{E}[N_1^{(n)} + \ldots + N_{X_n}^{(n)}] = \mathbb{E}[X_n]\mathbb{E}[N] = \mathbb{E}[N]^{n+1}$$ □

5.6 Optimal Stopping

Example 5.12

Suppose you play a game where you can roll a die up to three times. After each die, you can decide whether you want to stop and accept the current roll, or to continue rolling. What is the optimal stopping strategy?

Let $v(n,k)$ be the expected value of the game after you have rolled n times and the current state (or roll) is k. Working backward, $v(3,k) = k$ for any k. So the expected value of the third roll is 3.5.

Going back to the second roll, if you roll a 4, 5, or 6, stopping will give a greater payout than the expected payout of a third roll. So $v(2,4) = 4, v(2,5) = 5, v(2,6) = 6$. On the other hand, we should reroll on a 1, 2, or 3, so $v(2,1) = v(2,2) = v(2,3) = 3.5$. Thus the expected value of the second roll is 4.25.

Similarly, for the first roll, we would reroll on a 1, 2, 3, or 4, and stay on a 5 or 6. So $v(1,5) = 5, v(1,6) = 6$ and $v(1,1) = v(1,2) = v(1,3) = v(1,4) = 4.25$. So the expected value of the first roll (and thus the whole game) is $\frac{29}{6} = 4.83$.

We will now investigate a more general framework for stopping problems that will allow us to easily compute optimal strategies and expected values. First we will consider the time homogeneous case, where rewards are based only on the state, and invariant with respect to time.

Suppose $\mathcal{S}$ is the state space of a certain game, and suppose $f : \mathcal{S} \to \mathbb{R}$ is a function representing the reward for stopping at state i. We also have an initial distribution μ which contains the probabilities for starting in a given state. Lastly, let P be a stochastic matrix representing the probabilities for state changes.

As we can see, this setup has essentially determined a finite Markov chain. Note that we implicitly assumed the Markov property for state changes in this system. Then in order to solve this problem, we will need to compute the expected value of the game at each state (and time, if the game is time dependent). In other words we need to calculate $v(k)$ for each k, and $v(t,k)$ for each t, k.

Of course, at any state k, we have the option to stay or continue, and since we are seeking the optimal strategy, $v(k)$ would simply be the maximum of the two expected values. The expected value of staying is just the immediate reward $f(k)$. The value of continuing is the weighted values of the future states, $\sum_{i\in\mathcal{S}} p_{ki}v(i)$. So

$$v(k) = \max\{f(k), \sum_{i\in\mathcal{S}} p_{ki}v(i)\}$$

which we can also write in the form of the following two inequalities:

$$\begin{cases} v(k) \geq f(k) \\ v(k) \geq \sum_{i\in\mathcal{S}} p_{ki}v(i) \end{cases}$$

In the time dependent case, we will simply add in parameters for time in the value terms and adjust them appropriately:

$$\begin{cases} v(t,k) \geq f(k(\\ v(t,k) \geq \sum_{i\in\mathcal{S}} p_{ki}v(t+1,k) \end{cases}$$

If we have a maximum number of rounds T, then we add in the boundary condition $v(T,k) = f(k)$ (not an inequality).

Definition 5.22

The **expected value** of a game is the average payout, given by

$$\mathbb{E} = \begin{cases} \sum_{i \in \mathcal{S}} v(i)\mu_i \\ \sum_{i \in \mathcal{S}} v(1,i)\mu_i \end{cases}$$

Example 5.13

You roll a die and can stop at any point too receive the amount on the die. However, if you roll a 6 at any point, the game ends and you receive nothing. What is the optimal strategy and expected value?

First note that this game is time invariant. The transition matrix is

$$P = \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

At state 6, $v(6) = 0$ since we stop and get nothing.

At state 5, since the payout is never more than 5, $\sum_{i \in \mathcal{S}} p_{5i} v(i) \leq f(5) = 5$. So $v(5) = 5$.

At state 4, suppose for the sake of contradiction that it is optimal to continue. Then it is also optimal to continue for 1, 2, 3, so we would continue until a 5 or a 6. Each is equally likely, so the value would be 2.5. But this is lower than $f(4) = 4$, so it is optimal to stay.

At state 3, suppose continuing is optimal. Then we would also continue for 1, 2, meaning we continue until 4, 5, or 6. Then $v(1) = v(2) = v(3)$. Since we are assuming continuing is optimal, we also have

$$v(3) = \frac{1}{6}(v(3) * 3 + v(4) + v(5) + v(6)) = \frac{1}{2}v(3) + \frac{3}{2}$$

This implies that in the continuing case, $v(3) = \frac{9}{3} = 3$. This is precisely the value in the staying case as well, so we can take either strategy equally well.

Using the above discussion, we conclude that $v(1) = v(2) = 3$. Then the value of the game is equal to

$$\mathbb{E} = \frac{1}{6}(3 + 3 + 3 + 4 + 5) = 3$$