Year 2 — Introduction to Probability

Based on lectures by G. Farolfi
Discussions by A. Rajapakse
Notes taken by Aryan Dalal

MATH 170E, Summer Sessions 2025, UCLA

Catalog Description
Lecture, three hours; discussion, one hour. Requisites: courses 31A, 31B. Not open to students with credit for course 170A, Electrical and Computer Engineering 131A, or Statistics 100A. Introduction to probability theory with emphasis on topics relevant to applications. Topics include discrete (binomial, Poisson, etc.) and continuous (exponential, gamma, chi-square, normal) distributions, bivariate distributions, distributions of functions of random variables (including moment generating functions and central limit theorem). P/NP or letter grading.

Textbook Reading
Probability and Statistical Inference, *Hogg, Tanis, Zimmerman*

Contact
This document is a summary of the notes that I have taken during lectures at UCLA; please note that this lecture note will not necessarily coincide with what you might learn. If you find any errors, don't hesitate to reach out to me below:

Aryan Dalal, aryandalal@ucla.edu

Contents

0 Overview

What does it mean to know the chances of something to occur? When we say, "how likely is something going to happen", we are asking a question regarding *probabilities*. Most students certainly have some experience with probability in their day-to-day life. When you check the whether app for rain, you are often prompted a percentage regarding how likely it is for rain to occur around your surroundings. When you play traditional board games with a dice, you know that the probability that it lands on any one of its faces is one-sixths. We can ask more sophisticated questions that build on our basic idea of probability theory. "If I roll two die in monopoly, what are the chances the the sum of the two die is more than 6?" It sounds simple and in fact, it is; however, the idea of probability need not be confined to simply discrete events. Often, we also concern ourselves with events that occur on a continuum.

In our presentation of probability theory, we will give a thorough treatment to objects of both discrete and continuous nature. We begin laying the mathematical foundation for probabilities in Section 1 followed by introducing discrete random variables in Section 2. Continuous random variables will be introduced in Section 3; fundamental quantities/characteristics of random variables such as moments, spread, etc. will be discussed throughout all sections listed in this document.

1 Introduction to Probability Theory

1.1 Lecture 0: Sets and Probabilities

We begin with a thorough reading of *sets and probabilities*. Before we jump into the core concepts in probability, we want to understand the mathematical symbols and ideas that we will frequently refer to throughout this course. Naturally, this leads us to first give an appropriate treatment of *sets*. I have picked up the below text from an existing lecture note[1] for Math 170A: Probability Theory which requires real analysis as a pre-requisite.

In probability theory, a set represents some possible outcomes of some random process. For example, the set $\{1, 2, 3, 4, 5, 6\}$ has six elements which represent all possible rolls for a six-sided die. The set $\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$ has $6 \cdot 6 = 36$ elements, representing all possible ordered dice rolls for two-six sided dice. For example, the ordered pair $(2, 3)$ represents a roll of 2 on the first die, and a 3 on the second die. The set $[0, 1] \times [0, 1]$ in the plane $\mathbb{R}^2$ could represent the set of all possible locations of a dart thrown at a square dartboard. Eventually, we will assign probabilities to all elements of the set, but for now we will just focus on the sets themselves.

Definition 1.1 (Set, Element, Empty Set)**.** A *set* is a collection of objects. Each such object in the set is called an *element* of the set. If A is a set and x is an element of the set A, we write $x \in A$. If x is not an element of A, we write $x \notin A$. The set consisting of no elements is called the *empty set*, and this is denoted by $\varnothing$.

Definition 1.2 (Finite, Countably Infinite)**.** Let A be a set. We say that the set A is *finite* if there exists a nonnegative integer n such that A can be enumerated as a set of n elements. That is, we can write $A = \{x_1, x_2, \dots, x_n\}$. We say that the set A is *countably infinite* if A can be enumerated by the positive integers. That is, we can write $A = \{x_1, x_2, x_3, \dots\}$. We say that the set A is *uncountable* if: A is not finite, and A is not countably infinite.

Example 1.1. The set $\{1, 2, 3, 4, 5, 6\}$ is finite. The set $\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$ is finite. The set of even positive integers $\{2, 4, 6, 8, \dots\}$ is countably infinite. We could write the positive even integers in the following way:

$$\{2, 4, 6, 8, \dots\} = \{k \in \mathbb{R} : k/2 \text{ is a positive integer}\}$$

The last expression is read as "The set of all k in the set of real numbers such that $k/2$ is a positive integer."

The closed interval $[0, 1] = \{x \in \mathbb{R} : 0 \leq x \leq 1\}$ is uncountable. This (perhaps counterintuitive) fact is sometimes proven in Real Analysis, Math 131A. That is, there is no way to write $[0, 1]$ as a list $\{x_1, x_2, x_3, \dots\}$ where $x_i \in [0, 1]$ for every positive integer i.

Definition 1.3 (Subset)**.** Let A and B be sets. If every element of A is also an element of B, we say that A is a *subset* of B, and we write $A \subseteq B$ or $B \supseteq A$. If $B \subseteq A$ and $A \subseteq B$, we say that A and B are *equal* and write $A = B$.

[1] Heilman, S. (2016) *MATH 170A, Probability Theory, Winter 2016*, UCLA. www.stevenheilman.org/heilman/teach/170a.pdf

Example 1.2. We represent the roll of a single six-sided die by the universal set $\Omega = \{1, 2, 3, 4, 5, 6\}$. The set $A = \{1, 2, 3\}$ satisfies $A \subseteq \Omega$.

We can think of throwing darts at a flat, infinite board, so that the universal set is $\Omega = \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x, y) : x \in \mathbb{R} \text{ and } y \in \mathbb{R}\}$. We could imagine the dartboard itself as a square subset $[0, 1] \times [0, 1] \subseteq \Omega$. Or, perhaps, we could imagine a circular dartboard as a subset $\left\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\right\} \subseteq \Omega$.

Definition 1.4 (Complement)**.** Suppose A is a subset of some universal set Ω. The *complement of A in Ω* denoted A^c is the set $\{x \in \Omega : x \notin A\}$.

Example 1.3. If $\Omega = \{1, 2, 3, 4, 5, 6\}$ and if $A = \{1, 2, 3\}$, then $A^c = \{4, 5, 6\}$.

Remark. Note that we always have $\varnothing^c = \Omega$ and $\Omega^c = \varnothing$.

Definition 1.5 (Union, Intersection)**.** Let A, B be sets in soem universe Ω. The *union* of A and B, denoted $A \cup B$, is the set of elements that are in either A or B. That is,

$$A \cup B = \{x \in \Omega : x \in A \text{ or } x \in B\}$$

The *intersection* of A and B, denoted $A \cap B$, is the set of elements that are in both A and B. That is,

$$A \cap B = \{x \in \Omega : x \in A \text{ and } x \in B\}$$

The *set difference* of A and B, denoted $A \backslash B$, is the set of elements that are in A but not in B, that is,

$$A \backslash B = \{x \in A : x \notin B\}$$

Let n be a positive integer. Let $A_1, A_2, \ldots, A_n$ be sets in Ω. We denote

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n$$
$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n$$

Definition 1.6 (Countable Union, Countable Intersection)**.** Let $A_1, A_2, \ldots$ be sets in some universe Ω. The *countable union* of $A_1, A_2, \ldots$ denoted $\bigcup_{i=1}^{n} A_i$ is defined as follows.

$$\bigcup_{i=1}^{\infty} A_i = \{x \in \Omega : \exists \text{ a positive integer } j \text{ such that } x \in A_k\}$$

The *countable intersection* of $A_1, A_2, \ldots$ denoted $\bigcap_{i=1}^{\infty} A_i$ is defined as follows.

$$\bigcap_{i=1}^{\infty} A_i = \{x \in \Omega : x \in A_j, \forall \text{ positive integers } j\}$$

We can prove that the set of real numbers $\mathbb{R}$ can be written as the countable union

$$\mathbb{R} = \bigcup_{j=1}^{\infty} [-j, j]$$

We can do by showing containment from both sides. Likewise, we can also prove that the singleton set $\{0\}$ can be written as

$$\{0\} = \bigcap_{j=1}^{\infty} \left[\frac{-1}{j}, \frac{1}{j}\right]$$

Definition 1.7 (Disjointness, Partition)**.** Let n be a positive integer. Let A, B be sets in some universe Ω. We say that A and B are *disjoint* if $A \cap B = \varnothing$. A collection of sets $A_1, A_2, \ldots, A_n$ in Ω is said to be a *partition* of Ω if $\bigcup_{i=1}^{n} A_i = \Omega$, and if, $\forall i, j \in \{1, \ldots, n\}$ with $i \neq j$, we have $A_i \cap A_j = \varnothing$.

Proposition 1.1. Let A, B, C be sets in a universe Ω.

1. $A \cup B = B \cup A$
2. $A \cap (B \cup C) = (A \cap B) \cup (A \cup C)$
3. $(A^c)^c = A$
4. $A \cup \Omega = \Omega$
5. $A \cup (B \cup C) = (A \cup B) \cup C$
6. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
7. $A \cap A^c = \varnothing$
8. $A \cap \Omega = A$

Proposition 1.2 (De Morgan's Laws)**.** Let $A_1, A_2, \ldots$ be sets in some universe Ω. Then,

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c \qquad \left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c$$

Proof. We prove the first equality, since the second follows similarly. Suppose $x \in (\bigcup_{i=1}^{\infty} A_i)^c$. That is, $x \notin \bigcup_{i=1}^{\infty} A_i$. Recall that $\bigcup_{i=1}^{\infty} A_i = \{x \in \Omega : \exists j \in \mathbb{Z}^+ \text{ s.t. } x \in A_j\}$. Since x is not in the set $\bigcup_{i=1}^{\infty} A_i$, the negation of the definition of $\bigcup_{i=1}^{\infty} A_i$ applies to x. That is, x satisfies the negation of the statement: "$\exists$ a positive integer j such that $x \in A_j$". The negation of this statement is: "$\forall j \in \mathbb{Z}^+$, we have $x \notin A_j$". That is, $\forall$ positive integers j, we ahve $x \in A_j^c$. By definition of countable intersection, we conclude that $x \in \bigcap_{i=1}^{\infty} A_i^c$.

So, we have shown that $(\bigcup_{i=1}^{\infty} A_i)^c \subseteq \bigcap_{i=1}^{\infty} A_i^c$. To conclude, we must show that $(\bigcup_{i=1}^{\infty} A_i^c)$. So, let $x \in \bigcap_{i=1}^{\infty} A_i^c$. By reversing the above implications, we conclude that $x \in (\bigcup_{i=1}^{\infty} A_i)^c$. That is, $(\bigcup_{i=1}^{\infty} A_i)^c \supseteq \bigcap_{i=1}^{\infty} A_i^c$, and the proof is complete. □

We should now have a thorough understanding of the underlying mathematics and notation that we will use extensively throughout undergraduate probability theory. The above is a more rigorous introduction and is something that this course will not expect as the following is geared for students who have a primary interest in non-mathematical disciplines. That being said, I thought I would include this within the lecture notes because it is good to gain a thorough understanding prior to jumping into the core of the class. Also, as an undergraduate in mathematics, this was more of an authorial choice. I would like to thank Prof. Steven Heilman for his set of the section that I have used so far (the above is not my work and should not be assumed as such either). Prof. Heilman is an assistant professor at the University of Southern California.

1.2 Lecture 1: Basic Properties of Probabilities

When we roll a six-sided dice, there exists 6 possible outcomes

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

If the dice is *fair*, each outcome has a probability of 1/6 given by

$$\frac{1}{\#\text{ outcomes}}$$

Definition 1.8 (Outcome Space)**.** The set of all possible outcomes of a random experiment is called the *outcome space* denoted by S. The outcome space is also known as a sample space.

Suppose there exists several candies. We randomly select one candy from a box containing said candies with either strawberry or orange flavor. The outcome space S is given as

$$S = \{\text{strawberry},\ \text{orange}\}$$

Definition 1.9 (Event)**.** A subset $A \subseteq S$ is called an *event*

If $A_1, A_2, \ldots A_n \subseteq S$ satisfy $A_i \cap A_j = \varnothing, \forall i \neq j$, they are called *disjoint*. In other words, they're mutually exclusive. Likewise, if $A_1, A_2, \ldots, A_n \subseteq S$ satisfy $\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n = S$, then $\{A_i\}_{i=1,\ldots,n}$ are called *exhaustive*, that is, fully comprehensive.

Example 1.4. Let's roll a fair six-sided die. Some possible events for this random experiment are

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

Example 1.5. Consider two events A and B from the experiment of rolling a six-sided die

$$A := \{1, 2, 3\} \qquad\qquad B := \{4, 5, 6\}$$

Then,

$$A \cap B = \varnothing \implies A \text{ and } B \text{ are disjoint}$$
$$A \cup B = S \implies A \text{ and } B \text{ are fully comprehensive}$$

Definition 1.10 (Complement)**.** Given events $A_1, \ldots, A_k$, the *complement* of A_1 is

$$A_1^c = \bigcup_{i=2}^{k} A_i = A_2 \cup A_3 \cup \cdots \cup A_k$$

Definition 1.11 (Frequency, Relative Frequency)**.** Let n represent the number of times a random experiment is repeated. The *frequency* of an event A is the number of times the event actually occurred, and is denoted by $\mathrm{N}(A)$. The *relative frequency* of A in n experiments is

$$\frac{\mathrm{N}(A)}{n}$$

Suppose we roll a fair die ten times, $n = 10$, and we obtain $1, 1, 3, 5, 6, 3, 2, 3, 4, 4$. Let us consider the following events:

$$\begin{aligned} A &:= \{1\} \\ B &:= \{2, 4, 6\} \\ C &:= \{1, 2, 3, 4, 5, 6\} = S \\ D &:= \{\} = \varnothing \end{aligned}$$

Then,

$$\frac{\mathrm{N}(A)}{n} = \frac{2}{10} \qquad \frac{\mathrm{N}(b)}{n} = \frac{4}{10} \qquad \frac{\mathrm{N}(C)}{n} = \frac{10}{10} = 1 \qquad \frac{\mathrm{N}(D)}{n} = \frac{0}{10} = 0$$

Definition 1.12 (Probability)**.** The *probability* of an event A is the limit of $\frac{\mathrm{N}(A)}{n}$ as n grows larger and is denoted by the probability function $P(A)$.

Begin by noting that

$$0 \leq \frac{\mathrm{N}(A)}{n} \leq 1$$

Suppose we roll a fair six-sided die 6,000,000 times. We obtain the side 1 999,898 times. So, $A := \{1\} \implies P(A) = 1/6$. Observe that

$$\lim_{n \to 6000000} \frac{\mathrm{N}(A)}{n} = \frac{999,898}{6,000,000} \approx \frac{1}{6}$$

Definition 1.13 (Probability Function)**.** The *probability function*, P is defined as

$$P : S \longrightarrow [0, 1]$$

is a function that maps from the outcome space, S, to the interval $[0, 1]$. In other words, the function P is a function that assigns to each event $A \subseteq S$ a real number $P(A) \in [0, 1]$. Properties of P are given as

1. $P(A) \geq 0$ for any event A
2. $P(S) = 1$
3. Let $A_1, \ldots, A_k$ be disjoint such that $A_i \cap A_j = \varnothing \forall i \neq j$, then,

$$P(A_1 \cup \cdots \cup A_k) = P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} A_i = P(A_1) + \cdots + P(A_k)$$

Remark. 1. and 2. imply that $0 \leq P(A) \leq 1$

Theorem 1.1. *Given an event A and its complement A^c,*

$$P(A) = 1 - P(A^c)$$

Proof. By definition of mutual exclusion and the third property of $P(\ \cdot\ )$, we know that $P(A) + P(A^c) = P(A \cup A^c)$. Next, since A and A^c are complements of one another, we have $A \cup A^c = S$ so, $P(A \cup A^c) = P(S)$. By the second property of $P(\ \cdot\ )$, we have $P(S) = 1$ and thus,

$$P(A) + P(A^c) = 1 \iff P(A) = 1 - P(A^c)$$

□

Theorem 1.2. *For any outcome space S,*

$$P(\varnothing) = 0$$

Proof. Let $\varnothing = S^c$, then,

$$\begin{aligned} P(\varnothing) = P(S^c) &= 1 - P(S) && \text{By Theorem 1} \\ &= 1 - 1 && \text{By Property 2 of } P(\ \cdot\ ) \\ &= 0 \end{aligned}$$

□

Theorem 1.3. *Given two events A and B, if $A \subseteq B$, then*

$$P(A) \leq P(B)$$

Proof. Let $B - A := \{\text{elements of } B \text{ that are not in } A\}$. Then,

$$\begin{aligned} & 0 \leq P(B - A) && \text{By Property 1 of } P(\ \cdot\ ) \\ \iff & P(A) \leq P(A) + P(B - A) \\ & = P(A \cup (B - A)) && \text{By Definition of Mutual Exclusion} \\ & = P(B) && \text{Since } A \subseteq B \end{aligned} \tag{1}$$

Then, $P(A) \leq P(B)$ □

Theorem 1.4. *Given two events A and B*

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

Example 1.6. Suppose we want to elect a president and a vice president of a company choosing from the candidates Alice, Bob, Charlie, Dan. How many outcomes does this election have? No candidate can be both president and vice president.

Suppose the president is elected first:

1. 4 Candidates for President
2. 3 Candidates for Vice President

There are $4 \times 3 = 12$ possible outcomes.

Remark. The number of possible outcomes is the same if we elect the vice president first.

Theorem 1.5 (Multiplication Principle)**.** *Let experiment E_1 have n_1 outcomes and experiment E_2 have n_2 outcomes. Then, the composite experiment E_1E_2 has $n_1 \cdot n_2$ outcomes.*

Theorem 1.6 (Permutation)**.** *There are $n!$ ways to arrange n objects into a sequence, where*

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

Each arrangement is called a permutation of the object.

Example 1.7. How many ways can we select a president, vice president, CEO, CFO of a company from a group of 11 candidates?

1. 11 choices for President.
2. 10 choices for Vice President.
3. 9 choices for CEO.
4. 8 choices for CFO.

By multiplication principle, there are:

$$11 \cdot 10 \cdot 9 \cdot 8 = \frac{11!}{(11-4)!} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot \cancel{7} \cdot \cancel{6} \cdot \cancel{5} \cdot \cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}}{\cancel{7} \cdot \cancel{6} \cdot \cancel{5} \cdot \cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}}$$

Definition 1.14 (Sampling without Replacement)**.** Let r denote the elements we select from n different objects. Let this be an ordered sequence. If an object is not replaced after it has been selected from the sequence, then we are *sampling without replacement.*

Theorem 1.7. *There nPr ways to perform sampling without replacement of r objects from a group of size n, where*

$$nPr = \frac{n!}{(n-r)!} = n\,(n-1)\cdots(n-r+2)\,(n-r+1)$$

Example 1.8. A password is composed of 4 numerical single digits. How many possible passwords can we come up with? There exists 10 choices for all 4 digits. By multiplication principle, there are $10 \cdot 10 \cdot 10 \cdot 10 = 10^4$ possible outcomes.

Definition 1.15 (Sampling with Replacement)**.** Let r denote the elements we select from n different objects. Let this be an ordered sequence. If an object is replaced after it has been selected from the sequence, then we are *sampling with replacement.*

Theorem 1.8. *There are n^r ways to perform sampling with replacement of r objects from a group of size n.*

1.3 Lecture 2: Conditional Probability & Independence

Let's begin our discussion by an example. How many five-card hands can we form from a deck of 52 playing cards?

1.	52	choices for card 1
2.	51	choices for card 2
	⋮	
5.	48	choices for card 5

It's tempting to answer that the total number is ${}^{52}P_5 = 52!/47!$ but this is not the correct result. This is because the order of the cards in our hand does not matter. So, how many ways can we arrange the 5 cards in our hand? The result is 5!. As such, the correct answer is

$$\frac{{}^{52}P_5}{5!} = \frac{52!}{5!47!}$$

Definition 1.16 (Counting Without Order (Without Replacement))**.** Counting when order does not matter is when we select r elements from n different objects and the order of selection does not matter and the object is not replaced after it has been selected.

Theorem 1.9. *When the order does not matter, there are nCr ways to choose r elements from n different objects where*

$$nCr = \binom{n}{r} = \frac{nPr}{r!} = \frac{n!}{r!\,(n-r)!}$$

$\binom{n}{r}$ is also called the *binomial coefficient.*

Example 1.9. A basketball player takes 50 free throws and misses exactly 2. How many ways could they have done this?

$${}^{50}C_2 = \frac{50!}{2!48!} = \frac{50 \cdot 49}{2} = 1225$$

Theorem 1.10 (The Binomial Theorem)**.**

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

where

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

Example 1.10. We roll a fair die twice. Given that the first outcome is 3, what is the probability that the sum of the two dice rolls is 8? A non-rigorous answer is that if the first die is 3, the second must be 5. The probability of rolling a 5 on the second die is 1/6. A more rigorous version is as follows.

Let B be the event that we roll a 3 on the first roll. Then, possible outcomes are

$$B := \{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}$$

Let A be the event that the sum fo the dice is 8. Them, possible outcomes are

$$A := \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$$

It follows that $A \cap B = \{(3,5)\}$. The experiment of rolling a die twice has 36 outcomes (sampling with replacement) is given by $6^2 = 36$. Then,

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

and $P(B) = \frac{6}{36} = \frac{1}{6}$. Since B "occurred" already, B becomes our "new outcome space" for the experiment, and the probability of A occurring given that B occurred is

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

Definition 1.17 (Conditional Probability)**.** The probability that an event A occurs assuming that another event B has already happened is given by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0$$

This is called the *conditional probability* of A given B.

We can compute the probability of $P(A \cap B)$ in two ways,

$$P(A \cap B) = P(A) P(B \mid A), \qquad P(A) > 0$$

or

$$P(A \cap B) = P(B) P(A \mid B), \qquad P(B) > 0$$

It naturally follows that

$$P(A \mid B) = \frac{P(A)}{P(B)} P(B \mid A), \qquad P(A), P(B) > 0$$

Example 1.11 (Monty Hall Problem)**.** There exists 3 doors. One of the doors hides a car, the other two doors hide a goat. Note the following:

1. The participant wins whatever is hiding behind the door they choose.
2. A participant chooses one door, suppose door 1.
3. The host reveals that a goat hides behind a door that the participant didn't choose, suppose door 2.
4. The host asks the participant if they want to change their selected door, suppose from door 1 to door 3.

What's the best choice for the participant? Let

$$\begin{aligned} A &:= \{\text{car is behind door 1}\} \\ B &:= \{\text{door 2 is picked by host to reveal a goat}\} \\ C &:= \{\text{car is not behind door 1}\} \end{aligned}$$

Now, we don't know what $P(A \mid B)$ and $P(A^c \mid B)$ are, however, we do know that $P(B \mid A) = \frac{1}{2}$ because there are only two doors left that the host can pick to reveal a goat. It follows that $P(B \mid A^c) = \frac{1}{2}$. Now,

$$\begin{aligned} P(A \cap B) &= P(B \mid A)\, P(A) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \\ P(A^c \cap B) &= P(B \mid A^c)\, P(A^c) = \frac{1}{2} \cdot \frac{2}{3} = \frac{2}{6} \\ P(B) &= P(A \cap B) + P(A^c \cap B) = \frac{3}{6} \end{aligned}$$

We yield $P(B)$ because A and A^c are mutually exclusive and exhaustive. So, we can now calculate

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3} \\ P(A^c \mid B) &= 1 - P(A \mid B) = \frac{2}{3} \end{aligned}$$

We see that $P(A^c \mid B) > P(A \mid B)$ and so, the participant should switch doors.

Example 1.12. We flip a coin twice. Then,

$$S := \{(H,H), (H,T), (T,H), (T,T)\}$$

The coin is fair when

$$P((\ \cdot\ ,\ \cdot\ )) = \frac{1}{4}$$

Let

$$\begin{aligned} A &:= \{\text{heads on first flip}\} = \{(H,H), (H,T)\} \\ B &:= \{\text{tails on second flip}\} = \{(H,T), (T,T)\} \\ C &:= \{\text{tails on both flips}\} = \{(T,T)\} \end{aligned}$$

We have

$$P(B) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \qquad P(A) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

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

since $C \subset B$. "Knowledge" of C occurring changed the probability of B occurring. Indeed $P(B) \neq P(B \mid C)$ and

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{1/4}{2/4} = \frac{1}{2} = P(B)$$

"Knowledge" of A occurring did not change the probability of B. Note that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/2}{2/4} = \frac{1}{2} = P(A)$$

A and B are independent events.

Definition 1.18 (Independent Events)**.** Two events A and B are *independent* if

$$P(A \cap B) = P(A)\,P(B)$$

otherwise, A and B are called *dependent* events.

In particular, if A and B are independent and $P(B) > 0$:

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

Theorem 1.11. *If A and B are independent events, then so are*

1. *A and B^c*
2. *A^c and B*
3. *A^c and B^c*

Proof. We first prove (1), $P(A \cap B^c) = P(A)\,P(B^c)$. Now

$$\begin{aligned} P(A \cap B^c) &= P(A)\,P(B^c \mid A) && \text{by def. of cond. prob.} \\ &= P(A)\,[1 - P(B \mid A)] && \text{by def. of complement} \\ &= P(A)\,(1 - P(B)) && \text{by assumption of thm.} \\ &= P(A)\,P(B^c) \end{aligned}$$

Let's now prove (2), $P(A^c \cap B) = P(A^c)\,P(B)$. Begin with

$$\begin{aligned} P(A^c \cap B) &= P(B)\,P(A^c \mid B) && \text{by def. of cond. prob.} \\ &= P(B)\,[1 - P(A \mid B)] && \text{by def. of complement} \\ &= P(B)\,(1 - P(A)) && \text{by assumption of thm.} \\ &= P(B)\,P(A^c) \end{aligned}$$

We now prove (3), $P(A^c \cap B^c) = P(A^c)\,P(B^c)$. Now,

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

by Theorem from lecture 1, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. We have

$$\begin{aligned} P(A^c \cap B^c) &= 1 - P(A) - P(B) + P(A \cap B) \\ &= 1 - P(A) - P(B) + P(A)\,P(B) && \text{by assumption of thm.} \\ &= (1 - P(A))\,(1 - P(B)) \\ &= P(A^c)\,P(B^c) && \text{by def. of complement} \end{aligned}$$

□

Definition 1.19 (Mutual Independence)**.** $A_1, \ldots, A_k$ are pairwise independent if

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)\,P(A_2) \\ P(A_1 \cap A_3) &= P(A_1)\,P(A_3) \\ &\vdots \\ P(A_{k-1} \cap A_k) &= P(A_{k-1})\,P(A_k) \end{aligned}$$

$A_1, \ldots, A_k$ are mutually independent if they are pairwise independent and

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

Remark. Pairwise independence does not necessarily imply mutual independence.

Example 1.13. A fair die is rolled 6 tomes.

$$A_i := \left\{\text{side } i \text{ is rolled on } i^{\text{th}} \text{ roll}\right\}$$

Let's call A_i a "match".

$$P\left(A_i\right) = \frac{1}{6} \forall i = 1, \ldots, 6$$

$$P\left(A_i^c\right) = 1 - P\left(A_i\right) = 1 - \frac{1}{6} = \frac{5}{6}$$

Let $B := \{\text{at least one match happens}\}$. So, $B^c = \{\text{no matches occur}\}$. Note that $A_1, \ldots, A_6$ are mutually independent. Also, note that

$$B^c = \bigcap_{i=1}^{6} A_i$$

$$P\left(B^c\right) = P\left(\bigcap_{i=1}^{6} A_i\right) = \left(\frac{5}{6}\right)^6$$

Thus,

$$P\left(B\right) = 1 - P\left(B^c\right) = 1 - \left(\frac{5}{6}\right)^6$$

Example 1.14. We flip a fair coin twice. Let

$$A := \{(H,H), (H,T)\}$$

$$B := \{(H,H), (T,H)\}$$

$$C := \{(H,H), (T,T)\}$$

Because the coin is fair,

$$P\left((H,H)\right) = P\left((H,T)\right) = P\left((T,H)\right) = P\left((T,T)\right) = \frac{1}{4}$$

$$P\left(A\right) = \frac{1}{2} \qquad P\left(B\right) = \frac{1}{2} \qquad P\left(C\right) = \frac{1}{2}$$

A and B are independent if and only if $P\left(A \cap B\right) = P\left(A\right) P\left(B\right)$. Note that A, B and C are pairwise independent and this is verifiable.

$$P\left(A \cap B\right) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P\left(A\right) P\left(B\right)$$

$$P\left(A \cap C\right) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P\left(A\right) P\left(C\right)$$

$$P\left(B \cap C\right) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P\left(B\right) P\left(C\right)$$

We can also check if they're mutually independent. Now, A, B and C are mutually independent if and only if $P(A \cap B \cap C) = P(A) P(B) P(C)$. We know that

$$P(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(A) P(B) P(C)$$

So, we can conclude that A, B and C are not mutually independent.

Theorem 1.12 (Law of Total Probability)**.** *Let $B_1, \ldots, B_k$ be mutually exclusive and exhaustive events, and $P(B_i) > 0$ for $i = 1, \ldots, k$. Then*

$$P(A) = \sum_{i=1}^{k} P(A \mid B_i) P(B_i) = P(A \mid B_1) P(B_1) + \cdots + P(A \mid B_k) P(B_k)$$

Proof. Begin by

$$\begin{aligned} P(A) &= P\left(A \cap \left(\bigcap_{i=1}^{k} B_i\right)\right) = P(A \cap (B_1 \cap B_2 \cap \cdots \cap B_k)) \\ &= \sum_{i=1}^{k} P(A \cap B_i) = P(A \cap B_1) + \cdots + P(A \cap B_k) \\ &= \sum_{i=1}^{k} P(A \mid B_i) P(B_i) = P(A \mid B_1) P(B_1) + \cdots + P(A \mid B_k) P(B_k) \end{aligned}$$

□

Theorem 1.13 (Bayes' Theorem)**.** *Given $P(A), P(B) > 0$, then*

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

Remark. By Bayes' Theorem, we also have

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

Proof.

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} && \text{By definition of cond. prob.} \\ &= \frac{P(B \mid A) P(A)}{P(B)} && \text{By definition of cond. prob.} \end{aligned}$$

□

Bayes' Theorem is very nice. There's an entire field of 'Bayesian Statistics' after Bayes. Consider an example.

Example 1.15. Suppose we have 3 machines producing GPUs for computers. We have

1. Machine 1:

   (1) 1% of GPUs are defective
   (2) 25% of total GPUs produced
2. Machine 2:
   (1) 2% of GPUs are defective
   (2) 35% of total GPUs are produced
3. Machine 3:
   (1) 3% of GPUs are defective
   (2) 40% of total GPUs produced

Given a defective GPU, what are the probabilities that it was produced by machine 1, 2, or 3. Let

$$A := \{\text{GPU is defective}\}$$
$$B := \{\text{GPU was produced by machine } i, i = \{1, 2, 3\}\}$$

The question is asking $P(B_1 \mid A), P(B_2 \mid A), P(B_3 \mid A)$. We know $P(B_1) = 0.25, P(B_2) = 0.35, P(B_3) = 0.4$. By Bayes' Theorem,

$$P(B_1 \mid A) = \frac{P(A \mid B_1) P(B_1)}{P(A)}$$

We know $P(A \mid B_1) = 0.01$. However, we don't know $P(A)$. However, this is derivable from *Law of Total Probability* because B_1, B_2 and B_2 are mutually exclusive and exhaustive (since these GPUs can only be produced by one machine, not two or more of them at the same time).

$$P(A) = \sum_{i=1}^{3} P(A \mid B_i) = P(A \mid B_1) P(B_1) + P(A \mid B_2) P(B_2) + P(A \mid B_3) P(B_3)$$

Note that $P(A \mid B_1) = 0.01, P(A \mid B_2) = 0.02$ and $P(A \mid B_3) = 0.03$. So,

$$P(A) = 0.01 \cdot 0.25 + 0.02 \cdot 0.35 + 0.03 \cdot 0.4 = 0.0215$$

By Bayes' Theorem,

$$P(B_1 \mid A) = \frac{P(A \mid B_1) P(B_1)}{P(A)} = \frac{0.01 \cdot 0.25}{0.0215} = 0.1163$$
$$P(B_2 \mid A) = \frac{P(A \mid B_2) P(B_2)}{P(A)} = \frac{0.02 \cdot 0.35}{0.0215} = 0.3256$$
$$P(B_3 \mid A) = \frac{P(A \mid B_3) P(B_3)}{P(A)} = \frac{0.03 \cdot 0.40}{0.0215} = 0.5581$$

Given a defective GPU, there's a 11.62% chance it was produced by Machine 1, 32.56% chance it was produced by Machine 2 and 55.81% chance it was produced by Machine 3.

Example 1.16. Suppose we have three bowls.

1. B_1 = Bowl 1: 2 red chips, 4 white chips
2. B_2 = Bowl 2: 1 red chip, 2 white chips
3. B_3 = Bowl 3: 5 red chips, 4 white chips

One bowl is selected at random with the following probabilities:

$$P(B_1) = \frac{1}{3} \qquad P(B_2) = \frac{1}{6} \qquad P(B_3) = \frac{1}{2}$$

One chip is drawn at random from the selected bowl. Define

$$P(R) = \text{probability of drawing red}$$
$$P(W) = \text{probability of drawing white}$$

Observing that a red chip was drawn, what are the probabilities that it was drawn from B_1, B_2 or B_3. Note that B_1, B_2, B_3 are exhaustive ($P(B_1) + P(B_2) + P(B_3) = 1$) and are mutually exclusive, that is, only one bowl can be selected. We know that

$$P(R \mid B_1) = \frac{2}{6}$$
$$P(R \mid B_2) = \frac{1}{3}$$
$$P(R \mid B_3) = \frac{5}{9}$$

By Law of Total Probability,

$$P(R) = \sum_{i=1}^{3} P(R \mid B_i) P(B_i) = \frac{2}{6} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{3} + \frac{5}{9}\frac{1}{2} = \frac{5}{9}$$

By Bayes' Theorem,

$$P(B_1 \mid R) = \frac{P(R \mid B_1) P(B_1)}{P(R)} = \frac{1/3 \cdot 1/3}{4/9} = \frac{1}{4}$$
$$P(B_2 \mid R) = \frac{P(R \mid B_2) P(B_2)}{P(R)} = \frac{1/3 \cdot 1/6}{4/9} = \frac{1}{8}$$
$$P(B_3 \mid R) = \frac{P(R \mid B_3) P(B_3)}{P(R)} = \frac{5/9 \cdot 1/2}{4/9} = \frac{5}{8}$$

2 Discrete Random Variables

2.1 Lecture 3: Probability Mass and Cumulative Distribution Function

Definition 2.1 (Random Variable)**.** Given a random experiment with outcome space S, a *random variable* is a function $X : S \longrightarrow \mathbb{R}$ that assigns one real number $X_{(s)}$ to each $s \in S$.

Example 2.1. The outcome space of a die roll is $S := \{1, 2, 3, 4, 5, 6\}$. If we consider a random variable X representing this experiment, then a possible mapping for random variable X is

$$\begin{array}{lll} 1 \longmapsto 1 & 3 \longmapsto 3 & 5 \longmapsto 5 \\ 2 \longmapsto 2 & 4 \longmapsto 4 & 5 \longmapsto 6 \end{array}$$

Let's consider a less trivial example.

Example 2.2. Consider the outcome space of a coin flip:

$$S := \{H, T\}$$

Let X be a random variable representing this experiment. A possible mapping for X is

$$H \longmapsto 1 \qquad T \longmapsto 0$$

Another possible mapping for X is

$$H \longmapsto 0.35 \qquad T \longmapsto 0.99$$

Definition 2.2 (Discrete Random Variable)**.** A random variable X is *discrete* if the number of outcomes of X is countable.

Example 2.3. Let X = random number picked from $A := \{1, 2, 3\}$. Y= random real number picked from $B := [0, 1]$. Is X discrete? The answer is yes because the set A has a countable number of elements. Is Y discrete? The answer is no because the set B has an uncountable number of elements.

Let Z = random number picked from $C := \{1, 2, 3, \dots\}$ The answer is in fact yes. This is called a countable infinite set.

Definition 2.3 (Probability Mass Function)**.** The probability mass function (pmf) f of a discrete random variable X is given by

$$f(x) := P(X = x) \qquad \forall x \in \mathbb{R}$$

Remark. In probability theory notation, X is a random variable and x is the outcome that the random variable takes.

Example 2.4. Let's consider the experiment of measuring the volume of water that accumulates in a bucket of 2 liters that we keep in the backyard. The outcome space of this experiment is

$$S := [0, 2]$$

Let's consider the random variable X that applies the following mapping:

$$[0,1] \longmapsto 1 \qquad (1,2] \longmapsto 2$$

X is discrete but S has an uncountable number of elements.

Example 2.5. The probability mass function of X representing the experiment of rolling a fair die

$$f(1) = P(X=1) = \frac{1}{6} = f(2) = P(X=2) = \cdots = f(6) = P(X=6)$$

$$f(x) = P(X=x) = 0 \forall x \notin \{1,2,3,4,5,6\}$$

Definition 2.4 (Support of a Random Variable)**.** The *support* of a random variable X is the set of points for which $f(x) > 0$

Remark. For most discrete random variables, the support coincides with the outcome space.

Definition 2.5 (Cumulative Distribution Function)**.** The *cumulative distribution function* (cdf) $F(x)$ of a discrete random variable X is the function

$$F(x) = P(X \leq x), \qquad \forall x \in \mathbb{R}$$

Example 2.6. The cdf of X, $F(x)$ of a random variable representing the experiment of rolling a fair die:

$$F(x) = \begin{cases} 0 & \text{for } x < 1 \\ 1/6 & \text{for } x < 2 \\ 2/6 & \text{for } x < 3 \\ 3/6 & \text{for } x < 4 \\ 4/6 & \text{for } x < 5 \\ 5/6 & \text{for } x < 6 \\ 1 & \text{for } x \geq 6 \end{cases}$$

Let's consider $x = 1.5$, then ... to complete this

$$F(1.5) = P(X \leq 1.5) = P(X=1) = \frac{1}{6}$$

If we consider $x = 2.5$, then

$$F(2.5) = P(X \leq 2.5) = P(X=1) + P(X=2) = \frac{2}{6}$$

2.2 Lecture 4: Expectation & Variance

We will introduce the concept of expectation of a random variable. We will see that the expectation of a random variable a very specific case of a more general concept of a random variable called the *moment*. Let's do a quick review.

A random variable is a mapping $X : S \longrightarrow \mathbb{R}$. If X is discrete, its probability mass function is given as $f(\ \cdot\ ) : \mathbb{R} \longrightarrow [0, 1]$.

Example 2.7. Given an outcome space of the experiment of flipping a coin:

$$S := \{\text{Heads, Tails}\}$$

Let X be a discrete random variable such that

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

Then, the outcome space of X is

$$S_X = \left\{X_{(\text{Heads})}, Y_{(\text{Tails})}\right\} = \{0, 1\}$$

Assuming the coin is fair,

$$\begin{aligned} f(1) &= P(X = 1) = \frac{1}{2} \\ f(0) &= P(X = 0) = \frac{1}{2} \\ f(c) &= P(X = c) = 0 \qquad \forall c \notin S_X \end{aligned}$$

Example 2.8. We own a casino and we offer a payoff of x anytime a player rolls a number x on a fair, 6-sided die. How much should we charge each player to roll the die once to break-even in the long run.

We can calculate a weighted average of the payoffs, where the weights are the respective probabilities.

Let X be a random variable representing the single die roll,

$$\begin{aligned} P(X = 1) = \frac{1}{6} &\implies x = 1 \\ &\vdots \\ P(X = 6) = \frac{1}{6} &\implies x = 6 \end{aligned}$$

The weighted average is

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}.$$

The value 3.5 represents the expected earnings of a player playing the game. We can also think of it as the average earnings of a player that plays the game an infinite number of times. So, we must charge each player exactly 3.5 to break-even in the long run.

This weighted average is known as the *expectation* of a random variable.

Definition 2.6 (Expectation)**.** Let u be any function. The *expectation* (or *expected value*) of u for a random variable X of the discrete type is

$$\mathbb{E}\left[u\left(x\right)\right] = \sum_{x \in S_X} u\left(x\right) P\left(X = x\right)$$

where S_X is the outcome space of X.

Remark. The sum $\sum_{x \in S_X}$ can be an infinite sum. It's possible that it is not well-defined.

For a random variable X to have a well-defined expectation, we require the sum $\sum_{x \in S_X}$ to converge absolutely.

Example 2.9. Consider the previous example but now the payoff of each roll is x^2. The expected earnings are

$$\begin{aligned} u\left(X\right) &= X^2 \\ \mathbb{E}\left[X^2\right] &= \sum_{x \in S_X} u\left(x\right) P\left(X = x\right) = \sum_{x \in S_x} x^2 P\left(X = x\right) \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + \cdots + 6^2 \cdot \frac{1}{6} = \frac{91}{6} \end{aligned}$$

Example 2.10. Consider an experiment with only one outcome

$$S := \{\text{outcome}\}$$

Let X be a random variable such that

$$S_X := \left\{X_{(\text{outcome})}\right\} = \{c\}, \quad c \in \mathbb{R}$$

Then,

$$f\left(c\right) = P\left(X = c\right) = 1$$

by property of probability function.

Theorem 2.1 (Properties of Expectation)**.** *The expectation* $\mathbb{E}$ *satisfies the following properties,*

1. $\mathbb{E}\left[c\right] = c \quad \forall c \in \mathbb{R}$
2. $\mathbb{E}\left[cu\left(X\right)\right] = c\mathbb{E}\left[u\left(X\right)\right] \quad \forall c \in \mathbb{R}, u$ *a function.*
3. $\mathbb{E}\left[u_1\left(X\right) + u_2\left(X\right)\right] = \mathbb{E}\left[u_1\left(X\right)\right] + \mathbb{E}\left[u_2\left(X\right)\right]$, u_1, u_2 *are functions.*
4. $\mathbb{E}\left[X_1 X_2\right] = \mathbb{E}\left[X_1\right]\mathbb{E}\left[X_2\right]$ $\forall X_1, X_2$ *independent.*

Proof. (1) By definition of random variable, a constant is a random variable with $S_X := \{c\}$. Then,

$$\mathbb{E}\left[c\right] = \sum_{x \in S_X} xP\left(X = x\right) = c \cdot 1 = c$$

(2) By definition of expectation,

$$\mathbb{E}\left[cu\left(X\right)\right] = \sum_{x \in S_x} cu\left(X\right) P\left(X = x\right) = c \sum_{x \in S_X} u\left(X\right) P\left(X = x\right) = c\mathbb{E}\left[u\left(X\right)\right]$$

(3) By definition of expectation,

$$\begin{aligned}
\mathbb{E}\left[u_1\left(X\right) + u_2\left(X\right)\right] &= \sum_{x \in S_X} \left[u_1\left(X\right) P\left(X = x\right) + u_2\left(X\right) P\left(X = x\right)\right] \\
&= \sum_{x \in S_X} u_1\left(X\right) P\left(X = x\right) + \sum_{x \in S_X} u_2\left(X\right) P\left(X = x\right) \\
&= \mathbb{E}\left[u_1\left(X\right)\right] + \mathbb{E}\left[u_2\left(x\right)\right]
\end{aligned}$$

(4) By definition of independent discrete random variables (we will introduce a formal definition later in the course),

$$P\left(X_1 = x_1, X_2 = x_2\right) = P\left(X_1 = x_1\right) P\left(X_2 = x_2\right)$$

$$\begin{aligned}
\mathbb{E}\left[X_1 X_2\right] &= \sum_{x_1 \in S_{X_1}} \sum_{x_2 \in S_{X_2}} \left(x_1 \cdot x_2\right) P\left(X_1 = x_1, X_2 = x_2\right) \\
&= \sum_{x_1 \in S_{X_1}} \sum_{x_2 \in S_{X_2}} x_1 \cdot x_2 P\left(X_1 = x_1\right) P\left(X_2 = x_2\right) \\
&= \sum_{x_1 \in S_{X_1}} \sum_{x_2 \in S_{X_2}} x_1 P\left(X_1 = x_1\right) x_2 P\left(X_2 = x_2\right) \\
&= \left(\sum_{x_1 \in S_{X_1}} x_1 P\left(X_1 = x_1\right)\right) \left(\sum_{x_2 \in S_{X_2}} x_1 P\left(X_2 = x_2\right)\right) \\
&= \mathbb{E}\left[X_1\right] \mathbb{E}\left[X_2\right]
\end{aligned}$$

□

Example 2.11. Consider an experiment with two outcomes,

$$S := \{\text{Failure (F), Success (S)}\}$$

Let

$$p = \text{Probability of obtaining a success}$$

It follows that,

$$1 - p = \text{Probability of obtaining a failure}$$

The experiment is repeated until we obtain the first success and then we stop.

Let $X =$ number of iterations needed to obtain the first success. .

$$S_X := \{1, 2, 3, 4, 5, \dots\}$$

1 suggests that we need only 1 iteration to obtain a success; 5 suggests we need 5 iterations to obtain a success. X is discrete. Note that

$$f\left(1\right) = P\left(X = 1\right) = p$$

Assuming that all iterations of the experiment are independent from each other

$$\begin{aligned} f(2) &= P(X=2) = (1-p)\,p \\ f(3) &= P(X=3) = (1-p)(1-p)\,p = (1-p)^2\,p \\ &\vdots \\ f(n) &= P(X=n) = (1-p)^{n-1}\,p \end{aligned}$$

Definition 2.7 (Geometric Random Variable)**.** The *geometric random variable* X has

1. Support $= \mathbb{N}_{>0} = \{1, 2, 3, \dots\}$
2. Parameter $= p =$ (probability of obtaining a success in one attempt)
3. $P(X=n) = f(n) = (1-p)^{n-1}\,p,\ n \in \{1, 2, 3, \dots\}$

Theorem 2.2. *The cdf and expectation of a geometric random variable X with parameter p are*

1. *cdf* $= P(X \le n) = F(n) = 1 - (1-p)^n,\ n \in \{1, 2, 3, \dots\}$
2. $\mathbb{E}[X] = \sum_{n=1}^{\infty} n f(n) = \frac{1}{p}$

Proof. Let's begin with proving (2). Let $q = 1 - p$. Then,

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

By convergence of a geometric series

$$\begin{aligned} 1 + q + q^2 + q^3 + \cdots &= \frac{1}{1-q} \\ \frac{d}{dq}\left(1 + q + q^2 + q^3 + \cdots\right) &= \frac{d}{dq}\left(\frac{1}{1-q}\right) \\ \left(0 + 1 + 2q + 3q^2 + 4q^3 + \cdots\right) &= \frac{1}{(1-q)^2} \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{E}[X] &= \sum_{n=1}^{\infty} n q^{n-1} p = p\left(1 + 2q + 3q^2 + 4q^3 + \cdots\right) \\ &= p\frac{1}{(1-q)^2} = p\frac{1}{p^2} = \frac{1}{p} \end{aligned}$$

□

Proof of (1) as exercise.

Definition 2.8 (Mean, Variance, Standard Deviation)**.** Let X be a random variable. Then

1. Mean of X: $\mu := \mathbb{E}[X]$

2. Variance of X: $\sigma^2 := \text{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right]$
3. Standard Deviation of X: $\sigma := \sqrt{\text{Var}(X)}$

Remark. Variance and standard deviation are always non-negative.

Remark. Note that

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

by properties of expectation. Also, note that μ is a constant ($\mu = \mathbb{E}[X]$). Then, we can write the above as

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

Thus,

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

Which brings us to a second way of calculating variance,

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

This is often easier to use than the standard definition.

We have defined the above terms mathematically but what do they imply? Consider the following example.

Example 2.12. We're exploring investments. Let

1. X = we always get \$1
2. Y = we get \$2 with probability 0.5 and \$0 with probability 0.5
3. Z = we get \$10 with probability 0.9 and we lose \$80 with probability 0.1

We want to know what gives the highest return on average and which one is the riskiest to invest in.

$$\mathbb{E}[X] = \sum_{x \in S_X} x \cdot f(x) = 1 \cdot 1 = 1$$

$$\mathbb{E}[Y] = \sum_{y \in S_Y} y \cdot f(y) = 0 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2}$$

$$\mathbb{E}[Z] = \sum_{z \in S_Z} z \cdot f(z) = 10 \cdot \frac{9}{10} + (-80) \cdot \frac{1}{10} = 1$$

To answer our question on riskiest investment, we will use the variance. The variance tells us about the general distribution of the random variable. This will explored further soon.

$$\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \left(1^2 \cdot 1\right) - 1^2 = 0$$

$$\text{Var}(Y) = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 = \left(4 \cdot \frac{1}{2} + 0^2 \cdot \frac{1}{2}\right) - 1^2 = 1$$

$$\text{Var}(Z) = \mathbb{E}\left[Z^2\right] - (\mathbb{E}[Z])^2 = \left(10^2 \cdot \frac{9}{10} + (-80)^2 \cdot \frac{1}{10}\right) - 1^2 = 729$$

We conclude that investment Z is the most volatile.

2.3 Lecture 5: Moment Generating Function and The Bernoulli RV

Let's do a quick review from Lecture 4.

$$\mathbb{E}[X] = \sum_{x \in S_X} xf(x) = \text{ expected value of } X, \text{ mean of } X$$
$$\mathbb{E}\left[X^2\right] = \sum_{x \in S_X} x^2 f(x)$$

It turns out that expected values of different powers of random variables are called *moments* of a random variable or a distribution.

Definition 2.9 (Moment)**.** For any positive integer r, the rth *moment* of the random variable X is

$$\mathbb{E}[X^r]$$

Remark. The expected value of X, $\mathbb{E}[X]$, in addition being to the mean is also the 1st moment of X. Similarly, $\mathbb{E}\left[X^2\right]$ is the 2nd moment of X. Finally, $\mathbb{E}[X^r]$ is the rth moment of X.

Why are moments of a distribution important? They're important because they can be used to fully characterize a distribution. In other words, if two random distributions have exactly the same moments, that means that the two distributions are identical.

Before we define the significance of moments, let's consider an example of calculating moments.

Example 2.13. Let X be a random variable with outcome space defined by

$$S_X := \{-1, 0, 1\}$$

The probability mass function is given by

$$f(x) = \frac{1}{3} \forall x \in S_X$$

The 1st, 2nd, 3rd and 4th moments are

$$\mathbb{E}[X] = \sum_{x \in S_X} xf(x) = -1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0$$
$$\mathbb{E}\left[X^2\right] = \sum_{x \in S_X} x^2 f(x) = (-1)^2 \cdot \frac{1}{3} + (0)^2 \cdot \frac{1}{3} + (1)^2 \cdot \frac{1}{3} = \frac{2}{3}$$
$$\mathbb{E}\left[X^3\right] = \sum_{x \in S_X} x^3 f(x) = (-1)^3 \cdot \frac{1}{3} + (0)^3 \cdot \frac{1}{3} + (1)^3 \cdot \frac{1}{3} = 0$$
$$\mathbb{E}\left[X^4\right] = \sum_{x \in S_X} x^4 f(x) = (-1)^4 \cdot \frac{1}{3} + (0)^4 \cdot \frac{1}{3} + (1)^4 \cdot \frac{1}{3} = \frac{2}{3}$$

Moments can be used to verify whether two random variables have the same distributions (and if they do, they're identical distributions). In order to verify, we must check whether each distribution has well-defined moments (or the same number of well-defined moments).

Remark. It's possible that a random variable has no well-defined moments.

Example 2.14. If a discrete random variable X has outcome space

$$S_X := \mathbb{Z}^+$$

Then, the 1st moment involves taking the sum

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} xf(x)$$

If $\mathbb{E}[X]$ diverges, then X does not have a well-defined 1st moment. It's likely that if the 1st moment is not well-defined, then the rth moment is also not well-defined.

Remark. Moments are important because they they fully characterize a distribution of a random variable, and thus can be used to check if two random variables have the same distribution.

Example 2.15. Let X and Y be 2 discrete random variables. Suppose that both X and Y have well-defined moments upto the 3rd moment; these moments are identical. Moments of order higher than 3 are not well-defined.

$$\mathbb{E}[X] = \mathbb{E}[Y] \qquad \mathbb{E}\left[X^2\right] = \mathbb{E}\left[Y^2\right] \qquad \mathbb{E}\left[X^3\right] = \mathbb{E}\left[Y^3\right]$$

Then, X and Y have the same distribution.

Calculating the above is quite tedious. It's a routine computation and can get extremely long and complex if we are dealing with larger and more complicated probability mass functions. Does there exist a function that encodes all the moments of any distribution?

Definition 2.10 (Moment Generating Function)**.** The *moment generating function* of the discrete random variable X is

$$M(t) := \mathbb{E}\left[e^{tX}\right] = \sum_{x \in S_X} e^{tx} f(x)$$

Remark. t does not have any "meaning" or "interpretation" with respect to the distribution of X. It's only a parameter that's used to encode the moments of X into the moment generating function.

Remark. Note that by Taylor expansion of e^{tx} about 0, let $tx = y$, then

$$\begin{aligned} e^y &= e^0 + \frac{e^0}{1!}(y-0) + \frac{e^0}{2!}(y-0)^2 + \frac{e^0}{3!}(y-0)^3 + \cdots \\ &= \sum_{r=0}^{\infty} \frac{y^r}{r!} \\ e^{tx} &= \sum_{r=0}^{\infty} \frac{(tx)^r}{r!} \end{aligned}$$

Then, by definition of moment generating function,

$$\begin{aligned}\sum_{x\in S_X} e^{tx} f(x) &= \sum_{x\in S_X}\left(\sum_{r=0}^{\infty}\frac{t^r x^r}{r!}\right) f(x)\\ &= \sum_{r=0}^{\infty}\frac{t^r}{r!}\underbrace{\sum_{x\in S_X} x^r f(x)}_{r\text{th moment}}\\ &= 1 + t\mathbb{E}[X] + \frac{t^2}{2!}\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\mathbb{E}\left[X^3\right] + \cdots\end{aligned}$$

It's evident that this function explicitly encodes all moments of a random variable X. It follows that if two distributions have the same moment generating function, they're identical.

Theorem 2.3. *If two random variables have the same moment generating function, then they have the same support and probability mass function.*

Remark. By distribution of a discrete random variable, we mean the "shape" or functional form of the probability mass function of X, $f(x)$.

Example 2.16. Let X be a discrete random variable with

$$S_X := \{-1, 0, 1\}$$

$$f(x) = \begin{cases} 1/2 & x = -1\\ 1/6 & x = 0\\ 1/3 & x = 1\end{cases}$$

When a random variable has a continuous probability function, it will *not* be called a probability mass function, it will be called a probability density function. In this case, the random variable is called a continuous random variable. This is a quick preview of what's to come later in the course.

Theorem 2.4. *Let $r \in \mathbb{N}^+$. The rth moment of X is equal to the rth derivative of $M(t)$ with $t = 0$.*

$$\mathbb{E}[X^r] = \frac{d^r}{dt^r} M(t)\bigg|_{t=0}$$

Remark. This method calculate the rth moment of X works if and only if the rth derivative of $M(t)$ is well-defined around $t = 0$. If this method does not work (i.e., leads you to conclude that the rth moment of X is not well defined), it does not necessarily mean that X does not have a well defined rth moment. This conclusion (i.e., that X does not have a well-defined rth moment) can be reached only by calculating

$$\mathbb{E}[X^r] = \sum_{x\in S_X} x^r f(x)$$

If this sum is also not well-defined, then we can conclude that X does not have a well defined rth moment.

Example 2.17. Let X be a random variable with moment generating function

$$M(t) = \frac{pe^t}{1-qe^t} \qquad q = 1-p$$

We want to calculate the values of $\mathbb{E}[X]$ and $\mathbb{E}\left[X^2\right]$. We know that

$$\begin{aligned}
\mathbb{E}[X] &= \frac{d}{dt}M(t) \\
\frac{d}{dt}M(t) &= \frac{(pe^t)(1-qe^t)-(pe^t)(-qe^t)}{(1-qe^t)^2} = \frac{pe^t}{(1-qe^t)^2} \\
\mathbb{E}[X] &= \left.\frac{pe^t}{(1-qe^t)^2}\right|_{t=0} = \frac{pe^0}{(1-qe^0)^2} = \frac{p}{1-q^2} = \frac{p}{(1+p-1)^2} = \frac{1}{p} \\
\mathbb{E}\left[X^2\right] &= \left.\frac{d^2}{dt^2}M(t)\right|_{t=0} = \text{ as exercise}
\end{aligned}$$

Example 2.18. A basketball player shoots free-throws with probability of success p.

1. $1-p$ = probability of missing the free-throw.
2. X is a random variable for the experiment of shooting 1 free-throw.
3. $X = 0$, the free-throw is missed. $X = 1$, the free-throw is made.
4. $S_X := \{0, 1\}$
5. pmf $P(X = 0) = (1-p)$ and $P(X = 1) = p$.

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

Definition 2.11 (Bernoulli Random Variable)**.** A random variable X is a *Bernoulli* if it has

1. Support $:= \{0, 1\}$
2. Parameter p = probability of success
3. pmf $P(X = x) = f(x) = \begin{cases} p & ,= 1 \\ 1-p & x = 0 \end{cases}$

Example 2.19. We want to understand the mean and variance of a Bernoulli.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in S_X} xf(x) = 0 \cdot f(0) + 1 \cdot f(1) = 0 \cdot (1-p) + 1 \cdot p = p \\
\mathbb{E}\left[X^2\right] &= \sum_{x \in S_X} x^2 f(x) = 0^2(1-p) + 1^2 \cdot p = p \\
\sigma_X^2 &= \operatorname{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = p - p^2 = (1-p)p
\end{aligned}$$

Example 2.20. Suppose the same basketball player shoots 100 freethrows. Let

1. $Y = \#$ of successes out of 100 freethrows.
2. $p =$ probability of making 1 freethrow.

Assuming all freethrow attempts are independent.

$$\begin{aligned}
P(Y=0) &= (1-p)(1-p)(1-p)\cdots(1-p) = \prod_{i=1}^{100}(1-p) \\
&= (1-p)^{100} \\
P(Y=1) &= \sum_{i=1}^{100} P(\text{only } i\text{th shot is made}) \\
&= p(1-p)^{99} + (1-p)p(1-p)^{98} + \cdots + (1-p)^{99}p \\
&= 100p(1-p)^{99} \\
P(Y=2) &= \sum_{1\le i<j\le 100} P(\text{only } i\text{th and } j\text{th shots are made}) \\
&= pp(1-p)^{98} + p(1-p)p(1-p)^{97} + \cdots + (1-p)^{98}pp
\end{aligned}$$

How many arrangements of 2 objects out of 100 exist? It's ${}^{100}P_2$. Therefore,

$$P(Y=2) = \binom{100}{2} p^2 (1-p)^{98}$$

It follows that

$$P(Y=y) = \binom{100}{y} p^y (1-p)^{100-y} = f(y)$$

2.4 Lecture 6: The Negative Binomial and Poisson Random Variable

Example 2.21. Let $p =$ probability of obtaining tails when flipping a coin. We keep flipping the coin until we get 3 tails, and then we stop. We further assume that each coin flip is independent from each other. Let

$$X = \#\text{ of flips to obtain exactly 3 tails}$$

Then,

$$\begin{aligned} P(X=3) &= p \cdot p \cdot p = p^3 \\ P(X=4) &= (1-p)\,p^3 + p\,(1-p)\,p^2 + p^2\,(1-p)\,p = 3p^3\,(1-p) \\ &= \binom{3}{2} p^3\,(1-p) \end{aligned}$$

The last flip has to be tails. The first 3 flips have exactly 2 tails.

$$\begin{aligned} P(X=5) &= \left(1-p^2\right)p^3 + p\,(1-p)^2\,p^2 + p^2\,(1-p)^2\,p + \cdots + p\,(1-p)\,p\,(1-p)\,p \\ &= \binom{4}{2} p^3\,(1-p)^2 \end{aligned}$$

Once again, the last flip has to be tails. The first 4 flips have exactly 2 tails. If we were to generalize this, then

$$P(X=x) = \binom{x-1}{2} p^3\,(1-p)^{x-3}$$

This happens to be a very nice random variable called the Negative Binomial Random Variable

Definition 2.12 (Negative Binomial Random Variable)**.** A random variable X is a negative binomial if

1. Support $:= \{r, r+1, r+2, \dots\}$
2. Parameters $r =$ number of successes, $p =$ probability of success on 1 trial
3. pmf $f(x) = \binom{x-1}{r-1} p^r\,(1-p)^{x-r}$

Remark. Why is it called a negative binomial? Let $h(w) = (1-w)^{-r}$. It's possible to prove that

$$h(w) = (1-w)^{-r} = \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} w^k$$

which is called the *Negative Binomial Theorem.*

Theorem 2.5. *Let X be a negative binomial random variable with parameters r and p, ($X \sim nbin(r,p)$):*

1. $\mu_X = \dfrac{r}{p}$

2. $\sigma_x^2 = \dfrac{r\,(1-p)}{p^2}$

3. $M\,(t) = \dfrac{\left(pe^t\right)^r}{\left[1-(1-p)\,e^t\right]^r}$

Proof. By definition of the moment generating function,

$$\begin{aligned} M\,(t) &= \sum_{x=r}^{\infty} e^{tx} \binom{x-1}{r-1} p^r\,(1-p)^{x-r} \\ &= \frac{e^{tr}}{e^{tr}} \sum_{x=r}^{\infty} e^{tx} \binom{x-1}{r-1} p^r\,(1-p)^{x-r} \\ &= \left(pe^t\right)^r \sum_{x=r}^{\infty} e^{t(x-r)} \binom{x-1}{r-1} (1-p)^{x-r} \\ &= \left(pe^t\right)^r \sum_{x=r}^{\infty} \binom{x-1}{r-1} \left[(1-p)\,e^t\right]^{x-r} \end{aligned}$$

Remember that by the Negative Binomial Theorem

$$(1-w)^{-r} = \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} w^k$$

Let $x = r + k \iff k = x - r$. Let $w = (1-p)\,e^t$. It follows,

$$M\,(t) = \left(pe^t\right)^r \left(1-(1-p)\,e^t\right)^{-r}$$

□

We can use the proof of (3) to prove (1) and (2).

Remark. If $X \sim \text{nbin}\,(r,p)$, if we let $r = 1$, then $X \sim \text{geom}\,(p)$.

We can verify this.

$$f\,(x) = \binom{x-1}{r-1} p^r\,(1-p)^{x-r}\Big|_{r=1} = \binom{x-1}{1-1} p^1\,(1-p)^{x-1} = p\,(1-p)^{x-1}$$

Example 2.22. A 911 office receives an average of 7 calls per hour. What is the probability that the office receives exactly one call in the next hour? This has a non-trivial answer and we will need additional assumptions to answer this.

1. Split the next hour in 12 equal intervals of 5 minutes.
2. The expected number of calls in 1 hour = 7.
3. The expected number of calls in each 5-minute interval = 7/12. This is equivalent to assuming that the average 7 calls are randomly received throughout the hour.
4. p =probability of receiving a call in a 5-minute interval = 7/12.

5. Let $X = \#$ of phone calls received in 1 hour. Then

$$X \sim \text{bin}\left(n = 12, p = \frac{7}{12}\right)$$

Note that $\mathbb{E}[X] = np = 7$.

$$P(X = 1) = f(1) = \frac{12!}{1!\,(12-1)!}\left(\frac{7}{12}\right)^1\left(\frac{5}{12}\right)^{11} \approx 0.00046$$

Time is a continuous unit of measure, so it makes more sense to consider the case where we have an infinite number of time intervals within one hour, which corresponds to

$$\lim_{n\to\infty} X \sim \text{bin}\left(n, \frac{7}{n}\right) \iff \lim_{n\to\infty} f(x) = \text{ pmf of } X$$

Then,

$$\begin{aligned}
\lim_{n\to\infty} P(X = x) &= \lim_{n\to\infty} \frac{n!}{x!\,(n-x)!}\left(\frac{7}{n}]\right)^x\left(1 - \frac{7}{n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!}\frac{7^x}{n^x}\left(1 - \frac{7}{n}\right)^n\left(1 - \frac{7}{n}\right)^{-x} \\
&= \lim_{n\to\infty} \frac{7^x}{x!}\frac{n}{n}\frac{(n-1)}{n}\frac{(n-2)}{n}\cdots\frac{(n-x+1)}{n}\left(1 - \frac{7}{n}\right)^n\left(1 - \frac{7}{n}\right)^{-x}
\end{aligned}$$

Note that

$$\lim_{n\to\infty} \frac{n}{n} = \lim_{n\to\infty} \frac{n-1}{n} = \lim_{n\to\infty} \frac{n-2}{n} = \cdots = \lim_{n\to\infty} \frac{n-x+1}{n} = 1$$

This is a trivial result since we can divide the number by n.

$$\lim_{n\to\infty}\left(1 - \frac{7}{n}\right)^n = e^{-7}, \qquad \lim_{n\to\infty}\left(1 - \frac{7}{n}\right)^{-x} = 1$$

Definition 2.13 (Poisson Random Variable)**.** A random variable X is a *Poisson* if

1. $\text{Support}_X := \{0, 1, 2, \dots\}$
2. Parameters: $\lambda =$ average occurrence of an event per unit of measurement (rate)
3. $\text{pmf} = f(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$

Remark. The Poisson is the limit case of a random variable $X \sim \text{bin}\left(n, \frac{\lambda}{n}\right)$ when $n \to \infty$

Example 2.23. Flaws on a brand of magic tape occur once per 1200 feet on average. Let

$$Y = \#\text{ of flaws in a 1200-foot roll} \implies Y \sim \text{Poisson}(\lambda = 1)$$

$$X = \#\text{ of flaws in a 4800-foot roll of magic tape.}$$

We can expect that there will be, on average, 4 flaws per 4800-feet of magic tape. Then,

$$X \sim \text{Poisson}\left(4 \cdot \lambda = 4 \cdot 1 = 1\right)$$

What's the probability that there are no flaws in a 4800-foot roll?

$$P\left(X = 0\right) = \frac{4^0 e^{-4}}{0!} = 0.018$$

Theorem 2.6. *A Poisson randm variable X with rate λ has*

1. $\mu_X = \lambda$
2. $\sigma_X^2 = \lambda$
3. $M\left(t\right) = e^{\lambda\left(e^t - 1\right)}$

Proof. We prove (3). By the infinite order Taylor expansion of e^λ about 0,

$$\begin{aligned} e^\lambda &= 1 + \frac{e^0}{1!}\left(\lambda - 0\right) + \frac{e^0}{2!}\left(\lambda - 0\right)^2 + \frac{e^0}{3!}\left(\lambda - 0\right)^3 + \cdots \\ &= \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \end{aligned}$$

By definition of the moment generating function,

$$\begin{aligned} M\left(t\right) &= \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{e^{tx}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda} e^{\left(e^t \lambda\right)} = e^{\lambda\left(e^t - 1\right)} \end{aligned}$$

□

We can use the proof of (3) to prove (1) and (2).

Theorem 2.7. *Let X be the number of points from a random process, chosen from an interval of length 1. Assume that X is a Poisson random variable with parameter λ. Let Y be the number of points from the same random process, but chosen from an interval of length t. Then, Y is a Poisson random variable with parameter λt.*

Example 2.24. Let $X = \#$ of falling stars in the sky above Los Angeles in 1 second. We know that the average number of falling stars in the sky above LA is 60 per second. Then

$$\begin{aligned} &X \sim \text{Poisson}\left(\lambda = 60\right) \\ &Y = \#\text{ of falling stars in the sky above LA in 5 seconds} \end{aligned}$$

We want to calculate $P\left(Y \leq 2\right)$. First, we want to answer, every 5 seconds, how many falling stars are in the sky above LA on average? If the average is 60 per second, every 5 seconds, it's 300.

$$Y \sim \text{Poisson}\left(\lambda \cdot 5 = 60 \cdot 5 = 300\right)$$

Then,

$$\begin{aligned} P(Y \leq 2) &= P(Y=0) + P(Y=1) + P(Y=2) \\ &= \frac{300^0 e^{-300}}{0!} + \frac{300^1 e^{-300}}{1!} + \frac{300^2 e^{-300}}{2!} \end{aligned}$$

Example 2.25. A lightbulb company knows that 0.01% of its lightbulbs are defective. Let

$$X = \#\text{ of defective bulbs in a box of } 10000$$

What is the $P(X=3)$? We know that $X \sim \text{bin}(n=10000, p=0.0001)$

$$P(X=3) = \binom{10000}{3} (0.0001)^3 (0.9999)^{9997} \approx 0.061310$$

Let $\lambda = n \cdot p = 1$. Then, $Y \sim \text{Poisson}(\lambda = 1)$ and

$$P(Y=3) = 1^3 \frac{e^{-1}}{3!} \approx 0.061313$$

For n large and p small such that $n >> p$, the better the approximation of a binomial with a Poisson is.

Theorem 2.8. *Let X be a binomial random variable with parameters n and p, $X \sim bin(n,p)$. Let $np = \lambda$ and let Y be a Poisson random variable with parameter (mean rate) λ, $Y \sim Poisson(\lambda)$. For very large n and very small p, the probability*

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

can be approximated by

$$P(Y=x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

3 Continuous Random Variables

3.1 Lecture 7: Probability Density Function, Uniform and Exponential Random Variable

In this lecture, we will begin our discussion with *continuous random variables*. We won't stop using discrete random variables but our focus will be on continuous random variables.

We begin our discussion with a brief example. Remember that the cdf for a discrete Random Variable is a step function. Consider a Bernoulli random variable.

Example 3.1.

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

This implies that the cdf is a step function. Let X = real number selected at random (each number has an identical probability of being selected) from $[0, 1]$. Then,

$$P(X < 0) = 0$$
$$P(X \leq 1) = 1$$
$$P\left(X \leq \frac{1}{2}\right) = \frac{1}{2}$$

and

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

Definition 3.1 (Continuous Random Variable)**.** A random variable X is *continuous* if its cumulative distribution function (cdf), $F : \mathbb{R} \longrightarrow [0, 1]$, is *continuous*.

Remark. Remember that for discrete random variables, we have:

$$F(x) = P(X \leq x) = \sum_x P(X = x) = \sum_x f_{\text{pmf}}(x)$$

This is the relationship we have for cumulative distributive function and discrete random variables.

Definition 3.2 (Probability Density Function (pdf))**.** The *probability density function* of a continuous random variable X is the function $f : \mathbb{R} \longrightarrow \mathbb{R}$ given by

$$f(x) = \frac{d}{dx} F(x)$$

where $F(x)$ is the cumulative distribution function of X.

Example 3.2. Consider the random variable X from the previous example.

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases} \implies f(x) = \frac{d}{dx} F(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that the derivative is not defined in $x = 0, 1$. We can assume without loss of generality that $f(0) = 0$ and $f(1) = 0$. We can do so by thinking about the support of the random variable.

Theorem 3.1. *Let $f(x)$ be a probability density function. Then:*

1. $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$
3. $P(a < X < b) = \int_a^b f(x)\, dx \quad \forall a, b \in \mathbb{R}$
4. $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy \quad \forall x \in \mathbb{R}$

Remark.

1. A pdf is defined for continuous random variables, and is equal to
$$f(x) = \frac{d}{dx} F(x)$$
2. A pmf is defined for discrete random variables, and is equal to
$$f(x) = P(X = x)$$
3. For continuous random variables, $P(X = x)$ is always equal to 0.
4. For continuous random variables, in some instances, the pdf can be greater than 1.

Example 3.3. Let $a, b \in \mathbb{R}, a < b$. Let X = random real number chosen uniformly (each number has the same probability of being selected) from the interval $[a, b]$. Then,
$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$
$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$
Note that the support of X is the interval (a, b), because $f(x) > 0 \forall a < x < b$.

We note that the above example is indeed a very famous random variable called the *uniform random variable*.

Definition 3.3 (Uniform Random Variable)**.** A continuous random variable X is a *uniform* if

1. $\text{Support}_X := (a, b) \qquad \forall a, b \in \mathbb{R}, a < b$
2. pdf: $f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$

Remark. Note that

$$F(x) = \int_{-\infty}^{x} f(y)\, dy$$
$$= \begin{cases} \int_{-\infty}^{x} f(y)\, dy = \int_{-\infty}^{0} 0\, dy & x < a \\ \int_{-\infty}^{x} f(y)\, dy = \int_{-\infty}^{a} 0\, dy + \int_{a}^{x} \frac{1}{b-a}\, dy = \frac{x-a}{b-a} & a \le x \le b \\ \int_{-\infty}^{x} f(y)\, dy = \int_{-\infty}^{0} 0\, dy + \int_{a}^{b} \frac{1}{b-a}\, dy + \int_{b}^{x} dy = 1 & x > b \end{cases}$$

Example 3.4. Let X be a continuous random variable with pdf $f(x) = 4x$ and $\text{Support}_X := \left[0, \sqrt{1/2}\right]$. The cdf is

$$F(x) = \begin{cases} \int_{-\infty}^{x} f(y)\, dy = \int_{-\infty}^{x} 0\, dy = 0 & x < 0 \\ \int_{-\infty}^{x} f(y)\, dy = \int_{-\infty}^{0} dy + \int_{0}^{x} 4y\, dy = 2x^2 & 0 \le x \le \sqrt{1/2} \\ \int_{-\infty}^{x} f(y)\, dy = \int_{-\infty}^{0} 0\, dy + \int_{0}^{\sqrt{1/2}} 4y\, dy + \int_{\sqrt{1/2}}^{x} 0\, dy = 1 & x > \sqrt{1/2} \end{cases}$$

Theorem 3.2. *Let X be a continuous random variable with pdf $f(x)$. Then*

1. $\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx$
2. $\mathbb{E}[X^r] = \int_{-\infty}^{\infty} x^r f(x)\, dx$
3. $\sigma^2 = \mathbb{E}\left[X^2\right] - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2$
4. $M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$

Remark. If $M(t)$ is well-defined around $t = 0$, it still holds for continuous random variables that

$$\mathbb{E}[X^r] = \frac{d^r}{dt^r} M(t) \Big|_{t=0}$$

Theorem 3.3. *Let X be a uniform random variable on the interval $[a, b]$ such that $X \sim U(a, b)$. Then,*

1. $\mu_x = \dfrac{a+b}{2}$
2. $\sigma_x^2 = \dfrac{(b-a)^2}{12}$
3. $M(t) = \begin{cases} \dfrac{e^{tb} - e^{ta}}{t(b-a)} & t \neq 0 \\ 1 & t = 0 \end{cases}$

Proof. We begin with a proof of (3). For $t = 0$,

$$M(0) = \int_{-\infty}^{\infty} e^{0x} f(x)\, dx = \int_{-\infty}^{\infty} f(x)\, dx = 1$$

For $t \neq 0$,

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_a^b e^{tx} \frac{1}{b-a}\, dx \\ &= \frac{1}{b-a} \int_{-\infty}^{\infty} e^{tx}\, dx = \frac{1}{b-a} \left[\frac{e^{tx}}{t}\right]_a^b \\ &= \frac{e^{tb} - e^{ta}}{t(b-a)} \end{aligned}$$

□

We can prove (1) and (2) from (3).

Example 3.5. Ler X be a random variable with pdf $f(x) = |x|$ and $\text{Support}_X := (-1, 1)$. We compute the cdf of X as,

For $-1 < x \leq 0$,

$$\begin{aligned} F(x) = P(X \leq x) &= \int_{-1}^{x} |y|\, dy = \int_{-1}^{x} -y\, dy \\ &= \left[-\frac{y^2}{2}\right]_{-1}^{x} = -\frac{x^2}{2} + \frac{1}{2} \\ &= \frac{1-x^2}{2} \end{aligned}$$

For $0 < x < 1$,

$$\begin{aligned} F(x) = P(X \leq x) &= \int_{-1}^{x} |y|\, dy = \int_{-1}^{0} |y|\, dy + \int_0^x |y|\, dy \\ &= \int_{-1}^{0} -y\, dy + \int_0^x y\, dy \\ &= \left[-\frac{y^2}{2}\right]_{-1}^{0} + \left[\frac{y^2}{2}\right]_0^x \\ &= \frac{1}{2} + \frac{x^2}{2} = \frac{1+x^2}{2} \end{aligned}$$

Then,

$$F(x) = \begin{cases} 0 & x \leq -1 \\ \dfrac{1-x^2}{2} & -1 < x \leq 0 \\ \dfrac{1+x^2}{2} & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}$$

Suppose we want to find a real number from $(-1, 1)$, $\pi_{0.5}$, such that $P(X \leq \pi_{0.5}) = 0.5$.

$$F(\pi_{0.5}) = P(X \leq \pi_{0.5}) = \int_{-1}^{\pi_{0.5}} |x|\, dx = 0.5$$

Let's consider the case where $\pi_{0.5} \in (-1, 0]$, then

$$\frac{1-x^2}{2} = \frac{1}{2} \implies x = 0 = \pi_{0.5}$$

This allows us to understand the *percentile*.

Definition 3.4 (Percentile)**.** The $(100p)^{\text{th}}$ *percentile* is a number π_p such that

$$p = \int_{-\infty}^{\pi_p} f(x)\, dx = F(\pi_p) \qquad\qquad 0 \le p \le 1$$

Some important percentiles are

1. 50^{th} percentile = Median (or second quartile)
2. 25^{th} percentile = First Quartile
3. 75^{th} percentile = Third Quartile

Example 3.6. Let X represent the number of occurrences of a random event in an interval of time of length t. Let the random event have a mean rate of λt. Then,

$$X \sim \text{Poisson}(\lambda t) \qquad \text{and} \qquad f(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}$$

Let Y = time we wait before the first occurrence of the same random event. Then, probability of waiting more than t to observe the first occurrence is given by

$$P(Y > t) = P(X = 0)$$

that is the probability of observing 0 occurrences in interval of length t.

$$P(X = 0) = P(Y > t) = \frac{(\lambda t)^0}{0!} e^{-\lambda t}$$
$$P(Y \le t) = 1 - P(Y > t) = 1 - e^{-\lambda t} = F_Y(t) = \text{cdf of } Y$$

Y is a continuous random variable (time is a continuous measure, and its cdf is a continuous function) with

$$f_Y(t) = \frac{d}{dt} F_Y(t) = \lambda e^{-\lambda t}$$

$$\text{Support}_X := [0, \infty)$$

Definition 3.5 (Exponential Random Variable)**.** A continuous random variable X is an *exponential* if

1. $\text{Support}_x := [0, \infty)$
2. Parameter: λ = mean rate of underlying Poisson random variable
3. pdf: $f(x) = \lambda e^{-\lambda x}, \quad 0 \le x < \infty$

Remark. Another common reparametrization for exponential random variables is

$$\theta = \frac{1}{\lambda} \iff \lambda = \frac{1}{\theta} \iff f(x) = \frac{1}{\theta} e^{-\frac{1}{\theta} x}$$

3.2 Lecture 8: The Gamma & Chi-Square Random Variable

Let's consider an example where we can use the exponential random variable to model some random experiment.

Example 3.7. An iPad has a mean life span of 3 years. Let

$$X = \text{life span of Giulio's iPad}$$

Then, since θ represents the mean of an exponential random variable, we have

$$\theta = 3 \quad \text{and} \quad X \sim \exp(\theta)$$

What is the probability that Giulio's iPad breaks within a year after it is purchased?

$$P(x \leq 1) = \int_0^1 \frac{1}{3} \exp\left(-\frac{x}{3}\right) dx = \frac{1}{3}\left[-3\exp\left(-\frac{x}{3}\right)\right]_0^1 = 1 - e^{-1/3}$$

What is the probability that it will last for at least 3 years?

$$P(x \geq 3) = \int_3^\infty \frac{1}{3} \exp\left(-\frac{x}{3}\right) dx = \frac{1}{3}\left[-3\exp\left(-\frac{x}{3}\right)\right]_3^\infty = e^{-1}$$

Theorem 3.4. *Let X be an exponential random variable with parameter λ. Then,*

1. $\mu_x = \dfrac{1}{\lambda}$ *or* $\mu_x = \theta$ *where* $\theta = \dfrac{1}{\lambda}$
2. $\sigma_x^2 = \dfrac{1}{\lambda^2}$ *or* $\sigma_x^2 = \theta^2$ *where* $\theta = \dfrac{1}{\lambda}$
3. $M(t) = \dfrac{1}{1 - \frac{t}{\lambda}},\ t < \lambda$ *or* $M(t) = \dfrac{1}{1 - \theta t},\ t < \dfrac{1}{\theta}$

Proof. We'll start with proving 3.

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\theta} e^{-\frac{x}{\theta}}\, dx = \int_0^{\infty} e^{tx} \frac{1}{\theta} e^{-\frac{x}{\theta}}\, dx \\ &= \frac{1}{\theta}\int_0^{\infty} e^{tx} \frac{1}{\theta} e^{-\frac{x}{\theta}}\, dx = \frac{1}{\theta}\int_0^{\infty} e^{\left(t - \frac{1}{\theta}\right)x}\, dx = \frac{1}{\theta}\int_0^{\infty} e^{\left(\frac{t\theta - 1}{\theta}\right)x} \\ &= \frac{1}{\theta}\left[\frac{\theta}{t\theta - 1} e^{\left(\frac{t\theta - 1}{\theta}\right)x}\right]_0^{\infty} \end{aligned}$$

Note that for $t < \frac{1}{\theta}$, we have $\frac{t\theta - 1}{\theta} < 0$. Then, for $t < \frac{1}{\theta}$, we have

$$M(t) = \frac{1}{t\theta - 1}\left[0 - 1\right] = \frac{1}{1 - t\theta}$$

□

Suppose we let X =number of occurrences of some random event in a time interval of length t. Then, λt is the average number of random events per time period of length t. Let

$$\begin{aligned} Y_1 &= \text{time before 1st occurrence happens} \\ Y_2 &= \text{time between 1st and 2nd occurrences} \\ Y_3 &= \text{time between 2nd and 3rd occurrences} \end{aligned}$$

We have

$$P(X \geq 3) = \text{prob. that at least 3 occurrences happen in interval of length } t$$

Then,

$$P(Y_1 + Y_3 + Y_3 \leq t) = \text{prob. that we wait less than } t \text{ to observe the 3rd occurrence.}$$

We can model Y_1, Y_2, Y_3 as 3 independent exponential random variables with parameter λ:

$$\begin{aligned} Y_1 \sim \exp(\lambda) &\implies f_{Y_1}(t) = \lambda e^{-\lambda t} \\ Y_2 \sim \exp(\lambda) &\implies f_{Y_2}(t) = \lambda e^{-\lambda t} \\ Y_3 \sim \exp(\lambda) &\implies f_{Y_3}(t) = \lambda e^{-\lambda t} \end{aligned}$$

Now, let $Z = Y_1 + Y_2 + Y_3 =$ time until the 3rd occurrence. So,

$$\begin{aligned} P(Z \leq t) = P(X \geq 3) &= 1 - P(X < 3) \\ &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - e^{-\lambda t} - \lambda t e^{-\lambda t} - \frac{(\lambda t)^2}{2!} e^{-\lambda t} \\ &= 1 - \sum_{k=0}^{2} \frac{(\lambda t)^k e^{-\lambda t}}{k!} \end{aligned}$$

Let's generalize this to the case where Z =time until the wth occurrence. It follows that

$$P(X \geq w) = 1 - \sum_{k=1}^{w-1} \frac{(\lambda t)^k e^{-\lambda t}}{k!} \implies \text{cdf of } Z, \ (F_Z(w))$$

Then

$$f_Z(t) = \frac{d}{dt} F_Z(t) = \frac{\lambda (\lambda t)^{w-1}}{(w-1)!} e^{-\lambda t}$$

Definition 3.6 (Gamma Function)**.** The function $\Gamma(\alpha)$ is a *gamma* function if

$$\begin{aligned} \text{For } \alpha \in \mathbb{R}: &\quad \Gamma(\alpha) := \int_0^\infty y^{\alpha-1} e^{-y} \, dy \\ \text{For } \alpha \in \mathbb{N}_{>0}: &\quad \Gamma(\alpha) := (\alpha - 1)! \end{aligned}$$

Definition 3.7 (Gamma Random Variable)**.** The continuous random variable X is **gamma** if

1. $\text{Support}_X := [0, \infty)$
2. Parameters:

$$\alpha = \text{number of outcomes to be observed in a unit interval}$$
$$\theta = \text{mean of underlying exponential}$$

3. pdf: $f(x) = \dfrac{1}{\Gamma(\alpha)} \dfrac{x^{\alpha-1}}{\theta^{\alpha}} e^{-x/\theta}$

Remark. $f(x)$ in this definition can be derived from the pdf we derived in the previous example by setting $t = 1, w = \alpha, \lambda = \dfrac{x}{\theta}$

Example 3.8. Let the average lightsaber's life span be 3 years. Obi Wan replaces his light saver as soon as it breaks. Find the probability that Obi Wan will replace his light saver at least twice within six years.

$$X = \text{time until 2nd lightsaber breaks}$$

$$X \sim \text{Gamma}(\alpha = 2, \theta = 3)$$

Then

$$P(X \leq 6) = \int_0^6 \frac{1}{(2-1)!} \frac{x^{2-1}}{3^2} \exp\left(-\frac{x}{3}\right) = \frac{1}{9} \int_0^6 x \exp\left(-\frac{x}{3}\right) dx$$

By integration by parts, we yield

$$P(X \leq 6) = \frac{1}{9}\left[-3x \exp\left(-\frac{x}{3}\right) - 9 \exp\left(-\frac{x}{3}\right)\right]_0^6 = 1 - 3e^{-2}$$

Remark. If α is an integer, we can solve these type of problems (problems which can be solved using a Gamma random variable) using a Poisson.

Theorem 3.5. *Let X be a gamma random variable, then*

1. $\mu_x = \alpha\theta$
2. $\sigma_x^2 = \alpha\theta^2$
3. $M(t) = \dfrac{1}{(1-\theta t)^2}, \; t < \dfrac{1}{\theta}$

Proof. Once again, we only prove 3.

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_0^{\infty} e^{tx} \frac{1}{\Gamma(\alpha)} \frac{x^{\alpha-1}}{\theta^{\alpha}} e^{-x/\theta}\, dx \\ &= \frac{1}{\Gamma(\alpha)\theta^{a}} \int_0^{\infty} e^{tx} x^{\alpha-1} e^{-x/\theta}\, dx = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} \int_0^{\infty} x^{\alpha-1} \exp\left(-\left(\frac{1}{\theta} - t\right)x\right) dx \end{aligned}$$

Via an appropriate integration method, we yield

$$\begin{aligned} M(t) &= \frac{1}{\Gamma(\alpha)\theta^{\alpha}} \int_0^{\infty} \left(\frac{u}{1/\theta - t}\right)^{\alpha-1} e^{-u} \frac{du}{1/\theta - t} \\ &= \frac{1}{\Gamma(\alpha)\theta^{\alpha}(1/\theta - t)^{\alpha}} \int_0^{\infty} u^{\alpha-1} e^{-u}\, du \\ &= \frac{1}{\Gamma(\alpha)\theta^{\alpha}(1/\theta - t)^{\alpha}} \Gamma(\alpha) \\ &= \frac{(1/\theta)^{\alpha}}{(1/\theta - t)^{\alpha}} = \left(\frac{1/\theta - t}{1/\theta}\right)^{-\alpha} \\ &= (1 - t\theta)^{-\alpha} \end{aligned}$$

□

Let $X \sim \text{Gamma}\left(\alpha = \frac{r}{2}, \theta = 2\right)$. Then,

$$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right)} \frac{x^{r/2-1}}{2^{r/2}} e^{-x/2}$$

This special case of a Gamma random variable is called a Chi-Square random variable.

Definition 3.8 (Chi-Square Random Variable)**.** A continuous random variable X is *Chi-Square* if

1. $\text{Support}_X := [0, \infty)$
2. Parameters: r = degrees of freedom
3. pdf: $f(x) = \dfrac{1}{\Gamma(r/2)} \dfrac{x^{r/2-1}}{2^{r/2}} e^{-x/2}$

A Chi-Square random variable with r degrees of freedom is typically denoted by $\chi^2(r)$

3.3 Lecture 9: The Normal Random Variable

The normal random variable is a continuous random variable to which the binomial random variable with parameters n and $p = 1/2$ converges "in distribution" as $n \to \infty$.

$$X \sim \text{bin}\left(n, p = \frac{1}{2}\right)$$

Let $N\left(\mu, \sigma^2\right)$ = normal random variable with parameters μ, σ^2. Then,

$$X \xrightarrow{\text{d}} N\left(\mu, \sigma^2\right)$$

where "$\xrightarrow{\text{d}}$" means "converges in distribution".

Definition 3.9 (Normal Random Variable)**.** A continuous random variable X is *normal* if

1. $\text{Support}_X := (-\infty, \infty)$
2. Parameters:
$$\mu = \text{mean}$$
$$\sigma = \text{standard deviation}$$
3. pdf: $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{(x-\mu)^2}{2\sigma^2}\right]$

Theorem 3.6. *Let X be a normal random variable. Then,*

1. $\mu_x = \mu$
2. $\sigma_x^2 = \sigma^2$
3. $M(t) = \exp\left(\mu t + \dfrac{\sigma^2 t^2}{2}\right)$

Let X be a normal random variable with parameters $\mu, \sigma, X \sim N\left(\mu, \sigma^2\right)$. Suppose we want to transform this random variable into a normal with parameters $\mu = 0, \sigma^2 = 1$. We can do this by defining a new random variable:

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

Note that

$$\mu_Z = \mathbb{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{\mathbb{E}[X]}{\sigma} - \frac{\mu}{\sigma} = 0$$

$$\begin{aligned}
\mathbb{E}\left[Z^2\right] &= \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^2\right] = \mathbb{E}\left[\frac{X^2 - 2X\mu + \mu^2}{\sigma^2}\right] \\
&= \frac{1}{\sigma^2}\left[\mathbb{E}\left[X^2\right] - 2\mu\mathbb{E}[X] + \mu^2\right] = \frac{1}{\sigma^2}\left[\mathbb{E}\left[X^2\right] - 2\mu^2 + \mu^2\right] \\
&= \frac{1}{\sigma^2}\left[\mathbb{E}\left[X^2\right] - \mu^2\right] = \frac{1}{\sigma^2}\sigma^2 = 1
\end{aligned}$$

This process is known as standardization and Z is called a *standard normal random variable*:

$$Z \sim N(0, 1)$$

Definition 3.10 (Standard Normal Random Variable)**.** A continuous random variable Z is a *standard normal* if

1. $\text{Support}_Z := (-\infty, \infty)$
2. pdf: $f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$

such that $Z \sim N(0,1)$.

Let's make a few remarks.

Remark. Computing

$$P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy$$

is not possible using antiderivatives. Furthermore, standardization implies

$$P(X \leq x) = P\left(z \leq \frac{x-\mu}{\sigma}\right)$$
$$P(X \geq x) = 1 - P\left(z \leq \frac{x-\mu}{\sigma}\right)$$

where $X \sim N\left(\mu, \sigma^2\right), Z \sim N(0,1)$. Lastly, defining the cdf of a standard normal allows us to compute probabilities for any $X \sim N\left(\mu, \sigma^2\right)$.

Theorem 3.7. *Let Z be a standard normal random variable. Then,*

1. $\mu_z = 0$
2. $\sigma_z^2 = 1$
3. $M(t) = \exp\left(\frac{t^2}{2}\right)$

Example 3.9. Let $Z \sim N(0,1)$. We want to compute $P(Z \leq 1), P(-1 \leq Z \leq 1)$. Note that,

$$P(Z \leq 1) = \phi(1) = F_Z(1) = 0.8413$$

Now,

$$P(-1 \leq Z \leq 1) = P(Z \leq 1) - P(Z \leq -1) = \phi(1) - \phi(-1)$$

By symmetry of $f_Z(z)$, it must hold that

$$P(Z \leq -1) = P(Z \geq 1) = 1 - P(Z \leq 1) = 1 - \phi(1)$$

So,

$$\phi(-1) = 1 - \phi(1)$$

Then,

$$\begin{aligned} P(-1 \leq Z \leq 1) &= \phi(1) - \phi(-1) = \phi(1) - (1 - \phi(1)) \\ &= 2\phi(1) - 1 = 2(0.8413) - 1 = 0.6826 \end{aligned}$$

Theorem 3.8. *Let $X \sim N\left(\mu, \sigma^2\right)$. Then, the random variable*

$$V = \left(\frac{X-\mu}{\sigma}\right)^2 = Z^2 \sim \chi^2(1)$$

follows a Chi-Square distribution with 1 degree of freedom.

Example 3.10. Let $X \sim N\left(\mu = 1, \sigma^2 = 4\right)$. Then,

$$\begin{aligned}
P(X \leq 2) &= P\left(\frac{x-\mu}{\sigma} \leq \frac{2-\mu}{\sigma}\right) = P\left(z \leq \frac{2-1}{2}\right) \\
&= P\left(z \leq \frac{1}{2}\right) = \phi\left(\frac{1}{2}\right) = 0.6915 \\
P(1 \leq X \leq 3) &= P(X \leq 3) - P(X \leq 1) \\
&= P\left(\frac{x-1}{2} \leq \frac{3-1}{2}\right) - P\left(\frac{x-1}{2} \leq \frac{1-1}{2}\right) \\
&= P(Z \leq 1) - P(Z \leq 0) = \phi(1) - \phi(0) \\
&= 0.8413 - 0.5 = 0.3413
\end{aligned}$$

Example 3.11. Let $Z \sim N(0,1)$. We want to find a and b such that

$$\begin{aligned}
P(Z \leq a) = 0.9808 &\implies a \approx 2.07 \\
P(Z \geq b) = 1 - P(Z < b) &= 0.0024 \\
&\iff P(Z < b) = 0.9976 \implies b = 2.82
\end{aligned}$$

4 Bivariate Random Variables

4.1 Lecture 10: Joint Mass Function, Marginal Expectation & Variance

Suppose we want to first toss a fair coin, then we roll a fair die. Let $X =$ outcome of coin toss and $Y =$ outcome of die roll. We have

$$S_X = \{0, 1\} \qquad S_Y = \{1, 2, 3, 4, 5, 6\}$$

The *joint outcome space* of the two experiments represented by X and Y is the cartesian product of the two individual outcome spaces given by

$$\begin{aligned} S = S_X \times S_Y &= \{0, 1\} \times \{1, 2, 3, 4, 5, 6\} \\ &= \{(i, j) : 0 \leq i \leq 1, 1 \leq j \leq 6, i, j \in \mathbb{N}\} \end{aligned}$$

We can also define the *joint probability mass function*

$$f(x, y) = P(X = x, Y = y) \quad \forall (x, y) \in S_X \times S_y$$

Since both the coin and the die are fair by multiplication rile, we have that

$$f(x, y) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12} \quad \forall (x, y) \in S_X \times S_Y$$

Definition 4.1 (Joint Probability Mass Function, Joint Outcome Space)**.** Let X, Y be two discrete random variables. The *joint probability mass function* of X and Y is the function

$$f : S_X \times S_y \longrightarrow [0, 1], \quad f(x, y) = P(X = x, Y = y) \quad \forall x \in S_X, y \in S_Y$$

The *joint outcome space* is the cartesian product $S = S_X \times S_Y$.

Theorem 4.1. *Let X, Y be discrete random variables. The joint probability mass function f satisfies*

1. $0 \leq f(x, y) \leq 1$
2. $\sum_{(x,y)\in S} f(x, y) = 1$
3. *For every $A \subseteq S$, $P[(X, Y) \in A] = \sum_{(x,y)\in A} f(x, y)$*

Example 4.1. Let (X, Y) be the same bivariate discrete random variable defined in the previous example. We know that

$$f_X(0) = P(X = 0) = f_X(1) = P(X = 1) = \frac{1}{2}$$

$$f_Y(1) = P(Y = 1) = \cdots = f_Y(6) = P(Y = 6) = \frac{1}{6}$$

Note that $S_X \subset S, S_Y \subset S$. We can derive $f_X(x), f_Y(y)$ only knowing $f(x, y), S_X, S_Y$.

$$\begin{aligned} f_X(0) &= \sum_{y \in S_Y} f(0, y) = f(0, 1) + f(0, 2) + f(0, 3) + \cdots + f(0, 6) \\ &= 6 \cdot \frac{1}{12} = \frac{1}{2} \end{aligned}$$

$$f_Y(1) = f(x, 1) = f(0, 1) + f(1, 1) = \frac{1}{12} + \frac{1}{12} = \frac{1}{6}$$

$f_X(x), f_Y(y)$ are called *marginal pmfs* of X and Y, respectively.

Definition 4.2 (Marginal Probability Mass Function)**.** Let X, Y be discrete random variables with a joint probability mass function $f(x, y)$. The probability mass function of X is called the *marginal pmf of* X and is equal to

$$f_X(x) = P(X = x) = \sum_{y \in S_Y} f(x, y)$$

The probability mass function of Y is called the *marginal pmf of* Y and is equal to

$$f_Y(y) = P(Y = y) = \sum_{x \in S_X} f(x, y)$$

Remark. The marginal pmfs of X and Y are the pmfs of the random variables X and Y when considered independently.

Example 4.2. Let X, Y be two discrete random variables such that $X =$ drawing spades from a deck, $Y =$ drawing a king from a deck. It follows that,

$$S_X = \{0 = \text{not drawing spades}, 1 = \text{drawing spades}\}$$
$$S_Y = \{0 = \text{not drawing king}, 1 = \text{drawing king}\}$$

Furthermore, we can model

$$X \sim \text{Bernoulli}\left(p = \frac{13}{52}\right)$$
$$Y \sim \text{Bernoulli}\left(p = \frac{4}{52}\right)$$

We have

$$P(X = 1) = p = \frac{13}{52} = \frac{1}{4}$$
$$P(Y = 1) = \frac{4}{52} = \frac{1}{13}$$

Thus,

$$P(X = 1, Y = 1) = \frac{1}{52} = \frac{1}{4}\frac{4}{52} = P(X = 1) P(Y = 1)$$

X, Y are independent.

Definition 4.3 (Independent Random Variables)**.** Two discrete random variables X, Y are independent if, $\forall x \in S_X, y \in S_Y$, it holds that

$$P(X = x, Y = y) = P(X = x) P(Y = y)$$

Remark. Note that

$$P(X = x, Y = y) = f(x, y) = f_X(x) f_Y(y) = P(X = x) P(Y = y)$$

Example 4.3. Consider the following joint probability mass function

$$f(x, y) = \frac{xy^2}{13}$$

where $S_X = \{1, 2\}, S_Y = \{1, 2\}$ and $S = S_X \times S_Y$. Note that

$$\sum_{(x,y)\in S} f(x, y) > 1 \implies \text{violates the property of joint pmf.}$$

Is this enough to conclude that $f(x, y)$ is not a valid pmf? No! It depends on the joint support of the bivariate random variable! Let us consider

$$\text{Support} = \{(x, y) : x \leq y\} = \{(1, 1), (1, 2), (2, 2)\}$$

$S \neq$ Support. In particular, Support $\subset S$. This means that $f(2, 1) = P(X = 2, Y = 1) = 0$. In this case, $f(x, y)$ is a valid joint pmf because

$$f(1, 1) = \frac{1 \cdot 1^2}{13} = \frac{1}{13}$$
$$f(1, 2) = \frac{1 \cdot 2^2}{13} = \frac{4}{13}$$
$$f(2, 2) = \frac{2 \cdot 2^2}{13} = \frac{8}{13}$$

Which suggests that $f(1, 1) + f(1, 2) + f(2, 2) = 1$. Further note that

$$f_X(1) = f(1, 1) + f(1, 2) = \frac{1}{13} + \frac{4}{13} = \frac{5}{13}$$
$$f_Y(2) = f(1, 2) + f(2, 2) = \frac{4}{13} + \frac{8}{13} = \frac{12}{13}$$

Then,

$$f(1, 2) = \frac{4}{13} \neq \frac{5}{13} \cdot \frac{12}{13} = f_X(1) f_Y(2)$$

Thus, X and Y are not independent.

Example 4.4. Let $X =$ outcome of a fair coin toss (coin 1). $Y =$ outcome of a coin toss (coin 2) that is rigged such that always lands on the opposite outcome with respect to coin 1. So $S_X = \{0, 1\}, S_Y = \{0, 1\}$. Then, $S = S_X \times S_Y = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. We have Support $= \{(0, 1), (1, 0)\}$.

$$P(X = 0, Y = 0) = P(X = 1, Y = 1) = 0$$
$$P(X = 0, Y = 1) = P(X = 1, Y = 0) = \frac{1}{2}$$

X and Y are not independent because

$$P(X = 0) = P(X = 0, Y = 0) + P(X = 0, Y = 1) = 0 + \frac{1}{2} = \frac{1}{2}$$
$$P(Y = 0) = P(X = 0, Y = 0) + P(X = 1, Y = 0) = 0 + \frac{1}{2} = \frac{1}{2}$$

$$P(X = 0, Y = 0) = 0 \neq \frac{1}{2} \cdot \frac{1}{2} = P(X = 0) P(Y = 0)$$

Given $u(X, Y) = X + Y$, we can compute the expected value of $u(X, Y)$ as follows:

$$\mathbb{E}[u(X, Y)] = u(0, 0) f(0, 0) + \cdots + u(1, 1) f(1, 1) = \sum_{(x,y)\in S} u(x, y) f(x, y)$$
$$= (0 + 0) \cdot 0 + (0 + 1) \cdot \frac{1}{2} + (1 + 0) \cdot \frac{1}{2} + (1 + 1) \cdot 0 = 1$$

Definition 4.4 (Expectation of Bivariate Random Variable)**.** Let $u : S_X \times S_Y \longrightarrow (-\infty, \infty)$ be a function of two variables. Let X, Y be two discrete random variables with joint pmf $f(x, y)$. The expected value of $u(X, Y)$ is

$$\mathbb{E}[u(X,Y)] = \sum_{(x,y)\in S} u(x,y) f(x,y)$$

Given the former example, is it possible for us to evaluate $\mathbb{E}[X]$ knowing $f(x, y)$ and S_X, S_Y? Let $u(X, Y) = X$, then,

$$\begin{aligned}\mathbb{E}[u(X,Y)] &= \sum_{(x,y)\in S} xf(x,y) \\ &= 0 \cdot f(0,0) + 0 \cdot f(0,1) + 1 \cdot f(1,0) + 1 \cdot f(1,1) \\ &= 0 \cdot 0 + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} + 1 \cdot 0 = \frac{1}{2} \\ &= \mathbb{E}[X]\end{aligned}$$

Definition 4.5 (Marginal Expectation)**.** Let $u : S_X \times S_Y \longrightarrow (-\infty, \infty)$ be a function of two variables. Let X, Y be two discrete random variables with joint pmf $f(x, y)$. Let $u(X, Y) = X$, then,

$$\mu_X = \mathbb{E}[u(X,Y)] = \sum_{(x,y)\in S} xf(x,y)$$

Definition 4.6 (Marginal Variance)**.** Let X, Y be two discrete random variables with joint pmf $f(x, y)$. Let μ_X be the marginal expectation of X. Then, the marginal variance of X is

$$\sigma_X^2 = \sum_{(x,y)\in S} (x-\mu_X)^2 f(x,y) = \left(\sum_{(x,y)\in S} x^2 f(x,y) \right) - \mu_X^2$$

Example 4.5. In a class of 40 students, we can categorize the students based on their fantasy book preferences:

1. With probability = 0.45, a student prefers Lord of The Rings
2. With probability = 0.45, a student prefers Game of Thrones
3. With probability = 0.1, a student doesn't like fantasy books

We want to compute the probability that 18 students prefer Lord of The Rings, 16 prefer Game of Thrones, and 6 do not like fantasy books. We can define three "Binomial" random variables:

$$\begin{aligned}X &= \#\text{of students that like Lord of The Rings} \implies p_X = 0.45 \\ Y &= \#\text{of students that like Game of Thrones} \implies p_Y = 0.45 \\ Z &= \#\text{of students that do not like fantasy books} \implies p_Z = 0.1\end{aligned}$$

Note that $p_Z = 1 - P_X - P_Y$¿ Then, the probability that 18 students like Lord of The Rings, 16 students like Game of Thrones and 6 students do not like fantasy books Is

$$\begin{aligned}P(X = 18, Y = 16) = {} & p_X^{18} p_Y^{16} (1 - p_X - p_Y)^6 + p_X^{17} p_Y^{16} (1 - p_X - p_Y)^6 + \\ & p_X^{16} p_Y^{16} p_X^2 (1 - p_X - p_Y)^6 + \cdots + (1 - p_X - p_Y)^6 p_Y^{16} p_X^{18}\end{aligned}$$

The number of elements in the above sum is equivalent to

$$\frac{40!}{18!16!\left(40-18-16\right)!}$$

Then,

$$P\left(X=18,Y=16\right)=\frac{40!}{18!16!\left(40-18-18\right)!}p_X^{18}p_Y^{16}\left(1-p_X-p_Y\right)^6$$

Definition 4.7 (Trionomial Random Variable)**.** The two random variables X and Y form a *trinomial bivariate random variable* with parameters n, p_X, p_Y if

1.

$$\begin{aligned}S&=\{(x,y):0\leq x,y\leq n,x+y\leq n\}\\&=\{(0,0),\ldots,(0,n),(1,0),\ldots,(1,n-1),(2,0),\ldots,(n,0)\}\end{aligned}$$

2. pmf:

$$f\left(x,y\right)=P\left(X=x,Y=y\right)=\frac{n!}{x!y!\left(n-x-y\right)!}p_X^xp_Y^y\left(1-p_X-p_Y\right)^{n-x-y}$$

Remark. The trinomial random variable can be extended to the multinomial case, where the number of outcomes is >3.

4.2 Lecture 11: Covariance, Correlation, Conditional Variables

Given a random variable X,

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}\left[(X-\mathbb{E})[X]^2\right] = \mathbb{E}\left[(X-\mathbb{E}[X])(X-\mathbb{E})[X]\right] \\ &= \text{measure of "variability" of the random variable } X\end{aligned}$$

Given two random variables X, Y, a measure of "joint variability" is given by

$$\begin{aligned}\mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right] &= \mathbb{E}\left[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]\right] \\ &= \mathbb{E}[XY] - \mu_x\mu_y - \mu_x\mu_y + \mu_X\mu_y \\ &= \mathbb{E}[XY] - \mu_x\mu_y\end{aligned}$$

Definition 4.8 (Covariance)**.** Given two random variables X, Y, their *covariance* is defined as

$$\sigma_{xy} = \operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mu_x\mu_y$$

Remark. The covariance measures how X and Y tend to vary together. In other words, the covariance measures how much X and Y are dependent on each other. Note that

$$\begin{aligned}\text{X, Y Independent} &\implies \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] = \mu_x\mu_y \\ &\implies \operatorname{Cov}(X,Y) = 0\end{aligned}$$

Remark. The covariance is expressed in terms of two different unit of measure of X and Y. It also depends on the distributions of X and Y, making it not suitable for comparisons.

Definition 4.9 (Correlation Coefficient)**.** Given two random variables X, Y, the correlation coefficient of X, Y is

$$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Remark. The correlation coefficient ρ is always $-1 \le \rho \le 1$. If X, Y independent, it suggests $\rho = 0$.

Example 4.6. Let X, Y represent two independent fair coin tosses:

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{(x,y)\in S} xyf(x,y) \\ &= 0\cdot 0 f(0,0) + 0\cdot 1 f(0,1) + 1\cdot 0 f(1,0) + 1\cdot 1 f(1,1) \\ &= 0\cdot 0\cdot\frac{1}{4} + 0\cdot 1\cdot\frac{1}{4} + 1\cdot 0\cdot\frac{1}{4} 1\cdot 1\cdot\frac{1}{4} \\ &= \frac{1}{4}\end{aligned}$$

Note that

$$\mathbb{E}[X] = \mathbb{E}[Y] = 0\cdot\frac{1}{2} + 1\cdot\frac{1}{2} = \frac{1}{2}$$

$$\operatorname{Var}(X) = \operatorname{Var}(Y) = \left[0^2\cdot\frac{1}{2} + 1^2\cdot\frac{1}{2}\right] - \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

$$\sigma_X = \sigma_Y = \left(\frac{1}{4}\right)^{1/2} = \frac{1}{2}$$

It follows that $\sigma_{XY} = 0$ and $\rho = 0$. Now consider the random variables X and Y such that $f_X(x) = \frac{1}{3}$ for $x \in \{-1, 0, 1\}$ and $Y = X^2$. X and Y are clearly not independent because X fully determines Y. It's easy to show $\sigma_{XY} = 0$.

Example 4.7. Let X be the outcome of a fair coin toss and let $Y = X$. Then

$$\sigma_{XY} = \mathbb{E}[XY] - \mu_X \mu_Y = \mathbb{E}\left[X^2\right] - \mu_X^2 = \sigma_X^2$$
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sigma_X^2}{\sigma_X^2} = 1$$

Recall that

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

Example 4.8. We have a bag containing 3 coins. Each coin has different numbers on its sides:

1. Coin 1: Side A = 1, Side B = 2
2. Coin 2: Side A = 2, Side B = 3
3. Coin 3: Side A = 1, Side B = 3

We draw a coin at random. Let

$$X = \text{number on side } A$$
$$Y = \text{number on side } B$$

The joint pmf is given by

1. Coin 1: $(X = 1, Y = 2) \implies P(X = 1, Y = 2) = \frac{1}{3}$
2. Coin 2: $(X = 2, Y = 3) \implies P(X = 2, Y = 3) = \frac{1}{3}$
3. Coin 3: $(X = 1, Y = 3) \implies P(X = 1, Y = 3) = \frac{1}{3}$

with marginal pmfs given by $P(X = 1) = 2/3, P(X = 2) = 1/3, P(Y = 2) = 1/3$ and $P(Y = 3) = 2/3$. What's the probability that $Y = 2$ and $Y = 3$ given $X = 1$?

$$P(Y = 3 \mid X = 1) = \frac{P(X = 1, Y = 3)}{P(X = 1)} = \frac{1}{2}$$
$$P(Y = 2 \mid X = 1) = \frac{P(X = 1, Y = 2)}{P(X = 1)} = \frac{1}{2}$$

Note that $P(Y = y \mid X = 1) \neq P(Y = y)$ because X, Y are not independent.

Definition 4.10 (Conditional Probability for Bivariate Random Variables)**.** Let X, Y be discrete random variables with joint pmf $f(x, y)$. Then, the conditional probability of X given Y and Y given X are respectively

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x, y)}{f_Y(y)}$$
$$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f(x, y)}{f_X(x)}$$

Definition 4.11 (Conditional Discrete Random Variable)**.** Let X, Y be two discrete random variables. Given any element y in the support of Y (such that $f_Y(y) > 0$), the conditional random variable X given $Y = y$ ($X \mid Y = y$ or $X \mid y$) is characterized by the conditional pmf:

$$f_{X|y}(X \mid y) = \frac{f(x,y)}{f_Y(y)}$$

Continuing from the previous example, what are the mean and variance of $Y \mid X = 1$?

$$\mathbb{E}[Y \mid X = 1] = \sum_{y \in S_y} y f_{Y|X=1}(y \mid x = 1) = \frac{5}{2}$$

$$\mathbb{E}\left[Y^2 \mid X = 1\right] = \sum_{y \in S_y} y^2 f_{(Y|X=1)}(y \mid x = 1) = \frac{13}{2}$$

$$\sigma^2_{Y|X=1} = \mathbb{E}\left[Y^2 \mid X = 1\right] - \mu^2_{Y|X=1} = \frac{1}{4}$$

Definition 4.12 (Conditional Mean and Variance)**.** The *conditional mean* of $X \mid y$ is

$$\mu_{X|y} = \mathbb{E}[X \mid y] = \sum_{x \in S_X} x f_{X|y}(X \mid y)$$

The *conditional variance* of $X \mid y$ is

$$\sigma^2_{X|y} = \sum_{x \in S_X} \left(x - \mu_{X|y}\right)^2 f_{X|y}(X \mid y) = \mathbb{E}\left[X^2 \mid y\right] - \left(\mu_{X|y}\right)^2$$

Example 4.9. Let X, Y be two independent uniform random variables from the interval $[0,1]$. $X \sum U(0,1), Y \sim U(0,1)$. Note that

$$f_X(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad f_Y(y) = \begin{cases} 1 & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

because they are independent, we can find the joint pdf of X, Y by taking the product of the marginal pdfs:

$$f(x,y) = f_X(x) f_Y(y)$$

$$f(x,y) = \begin{cases} 1 & 0 \le x, y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Let $a, b \in [0,1]$. We want to compute $P(X \le a, Y \le b) =$ joint pdf of X, Y. We do so as

$$\begin{aligned} P(X \le a, Y \le b) = P(X \le a) P(Y \le b) &= \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)\, dx\, dy \\ &= \int_{-\infty}^{b} \int_{-\infty}^{a} f_X(x) f_Y(y)\, dx\, dy = \int_{-\infty}^{b} f_Y(y)\, dy \int_{-\infty}^{a} f_X(x)\, dx \\ &= \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)\, dx\, dx = \int_{0}^{b} \int_{0}^{a} 1\, dx\, dx = \int_{0}^{b} a\, dy = ab \end{aligned}$$

Definition 4.13 (Bivariate Continuous Random Variables)**.** Let X, Y be two continuous random variables. The joint probability density function $f(\ \cdot\ ,\ \cdot\ )$ of X and Y is the function such that

$$P(X \leq a, Y \leq b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)\ dx\, dy$$

Theorem 4.2. *The joint pdf $f(\ \cdot\ ,\ \cdot\ )$ of X and Y satisfies*

1. $f(x,y) \geq 0$
2. $f(x,y) =$ *when* $(x,y) \notin Support_{X,Y}$
3. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\ dx\, dy = 1$
4. *For any region A of the plane* $f(x,y)$

$$P[(X,Y) \in A] = \int_{A} f(x,y)\ dx\, dy$$

Remark. Remember that for bivariate discrete random variables with joint pmf $f(x,y)$ we have

$$f_X(x) = \sum_{y \in S_y} f(x,y) \qquad f_Y(y) = \sum_{x \in S_X} f(x,y)$$

For bivariate continuous random variables with joint pdf $f(x,y)$ we have that

$$f_X(x) = \int_{S_Y} f(x,y)\ dy \qquad f_Y(y) = \int_{S_X} f(x,y)\ dx$$

4.3 Lecture 12: The Conditional Standard Normal Random Variable

Given a joint pdf $f(x,y) = \frac{4}{3}(1-xy)$ of random variables X, Y, and a joint support given by $0 \leq x \leq 1, 0 \leq y \leq 1$, can we derive the marginal pdf of X? Remember that for bivariate discrete random variables, we have

$$f_X(x) = \sum_{y \in S_Y} f(x,y)$$

where $f(x,y)$ is the joint probability mass function. For bivariate continuous random variables, we have

$$f_X(x) = \int_{S_Y} f(x,y)\, dy$$

where $f(x,y)$ is the joint probability density function. It follows that

$$f_X(x) = \int_0^1 \frac{4}{3}(1-xy)\, dy = \left[\frac{4}{3}\left(y - \frac{xy^2}{2}\right)\right]_0^1 = \frac{4}{3}\left(1 - \frac{x}{2}\right)$$

for $0 \leq x \leq 1$. We can also in fact show that $\int_0^1 f_X(x)\, dx = 1$.

Definition 4.14 (Marginal pdf for Continuous Random Variables)**.** Let X, Y be two continuous random variables with joint pdf $f(x,y)$. The marginal pdf of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy \quad \forall x \in \text{Support}_X$$

The marginal pdf of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx \quad \forall y \in \text{Support}_Y$$

Now, given $f(x,y)$ from the start of this subsection, what can we say about $\mathbb{E}[X], \mathbb{E}[X^2]$ and σ_X^2?

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f(x,y)\, dy\right] dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\, dy\, dx = \int_0^1 \int_0^1 x \frac{4}{3}(1-xy)\, dy\, dx \\
&= \int_0^1 \left[\frac{4x}{3}\left(y - \frac{xy^2}{2}\right)\right]_0^1 dx = \int_0^1 \left[\frac{4}{3}\left(x - \frac{x^2}{2}\right)\right] dx \\
&= \left[\frac{4}{3}\left(\frac{x^2}{2} - \frac{x^3}{6}\right)\right]_0^1 = \frac{4}{9}
\end{aligned}$$

Definition 4.15 (Marginal Mean and Variance of Continuous Random Variables)**.** Let X, Y be two continuous random variables with joint pdf $f(x,y)$. The mean of X is given By

$$\mu_X = \mathbb{E}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\, dy\, dx$$

The variance of X is given by

$$\begin{aligned}\sigma_X^2 &= \mathbb{E}\left[(X-\mu_X)^2\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)^2 f(x,y)\,dx\,dx \\ &= \mathbb{E}\left[X^2\right] - (\mu_x)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^2 f(x,y)\,dx\,dx - (\mu_x)^2\end{aligned}$$

Example 4.10. Consider the joint pdf and joint support: $f(x,y) = 4$ for $0 \le x \le 1, 0 \le y \le \frac{x}{2}$. The support

$$\text{Support} = \{(x,y) \mid 0 \le x \le 1, 0 \le y \le x/2\}$$

is a triangle. Suppose we want to calculate $P(0 \le X \le 1, 0 \le Y \le 1/2)$, the probability over the whole joint support, then,

$$\begin{aligned}P(0 \le X \le 1, 0 \le Y \le 1/2) &= \int_0^1\int_0^{x/2} 4\,dy\,dx \\ &= \int_0^1 2x\,dx = 1\end{aligned}$$

Now suppose we want to derive the marginal pdfs of X and Y, then, by definition, we know

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dx \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

So,

$$\begin{aligned}f_X(x) &= \int_0^{x/2} f(x,y)\,dy = \int_0^{x/2} 4\,dy = 2x \\ f_Y(y) &= \int_{2y}^1 f(x,y)\,dx = \int_{2y}^1 4\,dx = 4(1-2y)\end{aligned}$$

for $0 \le x \le 1$ and $0 \le y \le 1/2$ respectively. We can further check that $f_X(x), f_Y(y)$ integrate to 1 over the marginal support, respectively. Note that

$$f_X(x) f_Y(y) = 2x(4-8y) = 8x - 16y \ne 4 = f(x,y)$$

Thus, X, Y are not independent.

Theorem 4.3. *If X, Y are two continuous random variables, they are independent if $\forall x \in S_X, y \in S_Y$,*

$$f(x,y) = f_X(x) f_Y(y)$$

Example 4.11. Let $f(x,y) = 1, 0 \le x \le 1, 0 \le y \le 1$. Then,

$$f_X(x) = \begin{cases}\int_0^1 1\,dy = 1 & 0 \le x \le 1 \\ \int_0^1 0\,dy = 0 & \text{otherwise}\end{cases} \qquad f_Y(y) = \begin{cases}\int_0^1 1\,dx = 1 & 0 \le y \le 1 \\ \int_0^1 0\,dx = 0 & \text{otherwise}\end{cases}$$

$\forall x \in [0,1], y \in [0,1]$, we have $f_X(x) f_Y(y) = 1 = f(x,y)$. Then, X, Y are independent.

Remark. Everything we learned about conditional random variables, for example, $X \mid Y = y$ or $Y \mid X = x$ applies to continuous random variables too, with the following distinction:

For discrete random variables,

$$f_{X|Y=y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}$$

where $f_{X|Y=y}(x \mid y), f(x,y), f_Y(y)$ are all pmfs.

For continuous random variables,

$$f_{X|Y=y}(x \mid y)$$

where $f_{X|Y=y}(x \mid y), f(x,y), f_Y(y)$ are all pdfs.

Recall that

$$X \sim N(0,1) \implies f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$
$$Y \sim N(0,1) \implies f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$$

Then, if X, Y are independent, $f(x,y) = f_X(x) f_Y(y) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right)$. What if X, Y are dependent? What is their joint pdf? Remember that

$$\rho = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

We can use ρ to define the joint pdf of two dependent standard normal random variables.

Theorem 4.4. *Let X, Y be two standard normal random variables with correlation coefficient ρ. The joint pdf of X and Y is given By*

$$f(x,y) = \frac{1}{2\pi\sqrt{1-p^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right)$$

Theorem 4.5. *Let X, Y be two standard normal random variables with correlation coefficient ρ. For any $x \in S_X$, the random variable $Y \mid x$ is a normal random variable with mean*

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

a linear function of x. And the variance of $Y \mid x$ is given by

$$\sigma^2_{Y|x} = 1 - \rho^2$$

which does not depend on x.

Example 4.12. Let X, Y be two standard normal random variables with $\rho = \frac{3}{5}$. Compute $P(0.83 < Y < 2.18)$ and $P(0 < Y \mid X = 5/3 < 2)$. We can do the first by

$$P(0.83 < Y < 2.18) = \phi(2.18) - \phi(0.83) = 0.1887$$

For the second, we can do the following:

$$P(0 < Y \mid X = 5/3 < 2) = P(0 < Y < 2 \mid X = 5/3)$$

By Theorem 4.5,

$$\mu_{Y|x} = \rho x = \rho\frac{5}{3} = 1$$
$$\sigma^2_{Y|x} = 1 - \rho^2 = \frac{16}{25}$$

For $Y \mid X = \frac{5}{3} \sim N\left(1, \frac{16}{25}\right)$, Let

$$Z = \frac{(Y \mid X = 5/3) - \mu_{Y|X}}{\sigma_{Y|X}} = \frac{(Y \mid X = 5/3) - 1}{4/5} \sim N(0,1)$$

Then,

$$\begin{aligned} P(0 < Y \mid X = 5/3 < 2) &= P\left(\frac{0-1}{4/5} < Z < \frac{2-1}{4/5}\right) = P(-1.25 < Z < 1.25) \\ &= \phi(1.25) - \phi(-1.25) = 2\phi(1.25) - 1 \\ &= 2(0.8944) - 1 = 0.7888 \end{aligned}$$

Recall that

$$X \sim N\left(\mu_X, \sigma^2_X\right) \implies f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left[-\frac{(x-\mu_X)^2}{2\sigma^2_X}\right]$$
$$Y \sim N\left(\mu_Y, \sigma^2_Y\right) \implies f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left[-\frac{(y-\mu_Y)^2}{2\sigma^2_Y}\right]$$

If X, Y are independent, then

$$f(x,y) = f_X(x) f_Y(y) = \frac{1}{2\pi\sigma_X\sigma_Y}\exp\left[-\frac{1}{2}\left(\left[\frac{x-\mu_X}{\sigma_X}\right]^2 + \left[\frac{y-\mu_Y}{\sigma_Y}\right]^2\right)\right]$$

What if X, Y are dependent? What is their joint pdf?

Theorem 4.6. *Let X, Y be normal random variables with correlation coefficient ρ. The joint pdf of X, Y, is given by*

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left[-\frac{q(x,y)}{2}\right]$$

where

$$q(x,y) = \frac{1}{1-\rho^2}\left(\left[\frac{x-\mu_X}{\sigma_X}\right]^2 - 2\rho\left[\frac{x-\mu_X}{\sigma_X}\right]\left[\frac{y-\mu_Y}{\sigma_Y}\right] + \left[\frac{y-\mu_Y}{\sigma_Y}\right]^2\right)$$

Theorem 4.7. *Let X, Y be two normal random variables with correlation coefficient ρ. For any $x \in S_X$, the random variable $Y \mid x$ is a normal random variable with mean*

$$\mathbb{E}[Y \mid x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

a linear of function of x and the variance of $Y \mid x$ is given by

$$\sigma^2_{Y|x} = \sigma^2_Y \left(1 - \rho^2\right)$$

which does not depend on x.

Remark. Note that by Theorem 4.7, we have

$$\mathbb{E}\left[X \mid y\right] = \mu_X + \rho \frac{\sigma_X}{\sigma_Y} \left(y - \mu_Y\right)$$

and

$$\sigma^2_{X|y} = \sigma^2_X \left(1 - \rho^2\right)$$

4.4 Lecture 13: Transformations of Random Variables & Convolution

We begin with a notation on remark for this lecture only. I will use the notation S_X to indicate the support of X, remember that in all other lectures, $S_X =$ outcome space of X.

Example 4.13. Let X be a discrete random variable with pmf and support $f_X(x) = P(X = x) = \frac{1}{3}, S_X := \{0, 1, 2\}$. Let us consider the random variable $Y = X^2$. We want to find the pmf and support of Y.

$$y = x^2 \forall x \in S_X \implies S_Y := \left\{0^2, 1^2, 2^2\right\} = \{0, 1, 4\}$$

$$f_Y(y) = \begin{cases} P(Y=0) = P\left(X^2 = 0\right) = P(X=0) = \frac{1}{3} \\ P(Y=1) = P\left(X^2 = 1\right) = P(X=1) = \frac{1}{3} \\ P(Y=4) = P\left(X^2 = 4\right) = P(X=2) = \frac{1}{3} \end{cases} = \begin{cases} \frac{1}{3} & y = 0 \\ \frac{1}{3} & y = 1 \\ \frac{1}{3} & y = 4 \end{cases}$$

Note that we are inverting the function $u(X) = X^2$ *only* on the support of X.

Recall that if $u : S_X \longrightarrow S_Y$ is an invertible function, its inverse is the function $v : S_Y \longrightarrow S_X$ such that

1. $v(u(x)) = x \quad \forall x \in S_X$
2. $u(v(y)) = y \quad \forall y \in S_Y$

Example 4.14. Let us consider a continuous random variable with support: $S_X := \{X \in \mathbb{R}\}$. Let $u : S_X \longrightarrow S_Y$ be $u(x) = x^2$ such that the random variable $Y = X^2$. Then, $v : S_Y \longrightarrow S_X$ is $v(y) = \sqrt{y}, y \geq 0$. Note that

$$v(u(x)) = v\left(x^2\right) = \sqrt{x^2} = |x| = x \quad \text{for } x \geq 0$$
$$u(v(y)) = u(\sqrt{y}) = (\sqrt{y})^2 = |y| = y \quad \text{for } y \geq 0$$

It might be tempting to say that u is invertible, but it is not because

$$\forall x < 0 : v(u(x)) = v\left(x^2\right) = \sqrt{x^2} = |x| = -x$$

This means that $u(x)$ is not invertible on the whole support of X.

Theorem 4.8. *Given a discrete random variable X with support $S_X = \{c_1, c_2, \dots\}$ and pmf f_X, and an invertible function $u : S_X \longrightarrow S_Y$, then the support and pmf of the random variable $Y = u(X)$ are:*

1. $S_Y = \{u(c_1, u(c_2), \dots)\}$
2. $f_Y(y) = f_X(v(y)) \quad \forall y \in S_Y$

where $v : S_Y \longrightarrow S_X$ is the inverse function of u.

Remark. We can still find S_Y and $f_Y(y)$ if u is not invertible on the whole S_X, but it is invertible on complementary partitions of S_X.

Example 4.15. Let X have $f_X(x) = \frac{1}{5}, S_X := \{-2,-1,0,1,2\}$. Let $Y = X^2$, we want to find $S_Y, f_Y(y)$. Note that X^2 is not invertible on the whole S_X, however, it is invertible on the two complementary partitions.

$$\begin{aligned} S_{X_1} &:= \{-2,-1\}, S_{X_2} := \{0,1,2\} \\ S_{Y_1} &:= \left\{(-2)^2, (-1)^2\right\}, S_{Y_2} := \left\{(0)^2, (1)^2, (2)^2\right\} \\ S_Y &:= S_{Y_1} \cup S_{Y_2} = \{0,1,4\} \end{aligned}$$

$$f_X(y) = \begin{cases} P(Y=0) = P\left(X^2=0\right) = P(X=0) = \frac{1}{5} \\ P(Y=1) = P\left(X^2=1\right) = P(X=-1) + P(X=1) = \frac{2}{5} \\ P(Y=4) = P\left(X^2=4\right) = P(X=-2) + P(X=2) = \frac{2}{5} \end{cases}$$

Example 4.16. Let $X \sim \text{Poisson}(\lambda = 4)$. Then $S_X := \{0,1,2,\dots\}, f_X(x) = \frac{4^x e^{-4}}{x!}$. Let $Y = u(X) = \sqrt{X}$. We want to find $S_Y, f_Y(y)$. $u(X)$ is invertible on the whole S_X, then, $S_Y := \left\{0,1,\sqrt{2},\sqrt{3},\dots\right\}$. $v : S_Y \longrightarrow S_X$ is $v(y) = y^2$.

$$\begin{aligned} f_Y(y) &= P(Y=y) = P(u(X)=y) = P(X=v(y)) \\ &= f_X(v(y)) = f_X\left(y^2\right) = \frac{4^{y^2} e^{-4}}{(y^2)!} \end{aligned}$$

Theorem 4.9. *Given a continuous random variable X with support $S_X = \{x : c_1 < x < c_2\}$ and pdf f_X and an invertible function $u : S_X \longrightarrow S_Y$, then the support and pdf of the random variable $Y = u(X)$ are:*

1. $S_Y = \{y : u(c_1) < y < u(c_2)\}$
2. $f_Y(y) = f_X(v(y)) |v'(y)|$

where $v : S_Y \longrightarrow S_X$ is the inverse functions of u.

Remark. We can still find $S_Y, f_Y(y)$ when u is not invertible on the whole S_X but it is invertible on complementary partitions of S_X.

Example 4.17. Let $X \sim \exp(\lambda = 1)$, such that $S_X := \{x \in \mathbb{R} : 0 \le x \le \infty\}$, $f_X(x) = e^{-x}$. Let $u : S_X \longrightarrow S_Y$ be $u(x) = e^X$ such that $Y = e^X$. We want to compute $S_Y, f_Y(y)$. Is e^x invertible on $0 < x < \infty$? As a matter of fact, yes. Then, by theorem 4.9,

$$\begin{aligned} S_Y :&= \{y = u(x) : x \in S_X\} = \left\{e^X : 0 < x < \infty\right\} \\ &= \left\{y : e^0 < y < e^\infty\right\} = \{y : 1 < y < \infty\} \end{aligned}$$

Note that $v : S_Y \longrightarrow S_X$ is $v(y) = \ln(y), y \in (1,\infty)$. The cdf of Y is

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(u(X) \le y) = P(X \le v(y)) \\ &= P(X \le \ln(y)) = \int_0^{\ln(y)} e^{-x}\,dx \\ f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} \int_0^{\ln(y)} e^{-x}\,dx \\ &= \frac{d}{dy} (\ln(y))\, e^{-\ln(y)} = \frac{1}{y} \frac{1}{e^{\ln(y)}} = \frac{1}{y^2} \end{aligned}$$

Example 4.18. Let X be a continuous random variable with support $S_X := \{x \in \mathbb{R} : -1 < x < 2\}$, $f_X(x) = \frac{x^2}{3}$. Let $u : S_X \longrightarrow S_Y$ be $u(x) = x^2$. We want to compute $S_Y, f_Y(y)$. Is $u(x)$ invertible on the whole S_X? The answer is no. However, $u(x)$ is invertible on

$$S_{X_1} := \{x \in \mathbb{R} : -1 \leq x \leq 0\} \text{ and } S_{X_2} := \{x \in \mathbb{R} : 0 < x < 2\}$$

From Theorem 4.9 on the two partitions S_{Y_1} and S_{Y_2}, we have

$$\begin{aligned} S_{Y_1} := \{u(x) : x \in S_{X_1}\} = \left\{x^2 : -1 \leq x \leq 0\right\} \\ = \{y \in \mathbb{R} : 0 \leq y < 1\} \\ S_{Y_2} := \{u(x) : x \in S_{X_2}\} = \{y \in \mathbb{R} : 0 < y < 4\} \end{aligned}$$

Then, $S_Y := S_{Y_1} \cup S_{Y_2} = \{y \in \mathbb{R} : 0 < y < 4\}$. Now, u is invertible on both S_{X_1} and $S_{X_@}$, So

$$\begin{aligned} v_1 : S_{Y_1} \longrightarrow S_{X_1} \text{ is } v_1(y) = -\sqrt{y} \text{ for } 0 \leq y < 1, v_1'(y) = -\frac{1}{2\sqrt{y}} \\ v_2 : S_{Y_2} \longrightarrow S_{X_2} \text{ is } v_2(y) = \sqrt{y} \text{ for } 0 < y < 4, v_2'(y) = \frac{1}{2\sqrt{y}} \end{aligned}$$

We apply Theorem 4.9 on separate partitions to yield

$$\begin{aligned} f_{Y_1}(y) = f_X(v_1(y) \mid v_1'(y)) = \frac{\left(-\sqrt{y}\right)^2}{3}\left|-\frac{1}{2\sqrt{y}}\right| = \frac{\sqrt{y}}{6} \text{ for } 0 \leq y \leq 1 \\ f_{Y_2}(y) = f_X(v_2(y) \mid v_2'(y)) = \frac{\left(\sqrt{y}\right)^2}{3}\left|\frac{1}{2\sqrt{y}}\right| = \frac{\sqrt{y}}{6} \text{ for } 0 < y < 4 \end{aligned}$$

Then, $f_Y(y) = f_{Y_1}(y) + f_{Y_2}(y)$.

$$f_Y(y) = \begin{cases} f_{Y_1}(y) + f_{Y_2}(y) = \frac{\sqrt{y}}{6} + 0 = \frac{\sqrt{y}}{6} & y = 0 \\ f_{Y_1}(y) + f_{Y_2}(y) = \frac{\sqrt{y}}{6} + \frac{\sqrt{y}}{6} = \frac{\sqrt{y}}{3} & 0 < y < 1 \\ f_{Y_1}(y) + f_{Y_2}(y) = 0 + \frac{\sqrt{y}}{6} = \frac{\sqrt{y}}{6} & 1 \leq y < 4 \\ 0 & \text{otherwise} \end{cases}$$

Example 4.19. Let $X \sim U(1,2), Y \sim U(-2,-1)$. Let us consider the random variable $Z = X + Y$, and assume that $X < Y$ are independent. We want to find hte pdf of Z.

$$F_Z(a) = P(Z \leq a) = P(X + Y \leq a) = P(X \leq a - Y) = F_X(a - y)$$

We cannout compute $F_X(a - Y)$ directly because Y is a random variable. However, if X, Y are independent, we can integrate over all possible values of y to express $F_X(a - Y)$.

$$F_{X+Y}(a) = F_X(a - Y) = \int_{-\infty}^{\infty} F_X(a - y) f_Y(y)\, dy$$

Then, by definition of pdf,

$$\begin{aligned} f_{X+Y}(a) &= \frac{d}{da}\int_{-\infty}^{\infty} F_X(a - y) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \frac{d}{da} F_X(a - y) f_Y(y)\, dy = \int_{-\infty}^{\infty} f_X(a - y) f_Y(y)\, dy \end{aligned}$$

Then, in our example

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy$$
$$f_X(z-y) = \begin{cases} 1 & 1 \le z-y \le 2 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & z-2 \le y \le z-1 \\ 0 & \text{otherwise} \end{cases}$$
$$f_Y(y) = \begin{cases} 1 & -2 \le y \le -1 \\ 0 & \text{otherwise} \end{cases}$$

Now, since $f_X(z-y)$ and $f_Y(y)$ have different bounds, we have to compute 2 integrals. Consider the first case where $-1 \le z \le 0$. In this interval, $-2 \le z \le z-1$.

$$\int_{-2}^{z-1} f_X(a-y) f_Y(y)\, dy = \int_{-} 2^{z-1} 1\, dy = z-1+2 = z+1$$

Now, consider the second case where $0 < zleq1$. In this interval, $z-2 \le y \le -1$.

$$\int_{z-2}^{-1} f_X(a-y) f_Y(y) = \int_{z-2}^{-1} 1\, dy = -1+2-z = 1-z$$

Then,

$$f_Z(z) = \begin{cases} z+1 & -1 \le z \le 0 \\ 1-z & 0 < z \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Definition 4.16 (Convolution)**.** Let X and Y be two independent continuous random variables having pdfs $f_X(x), f_Y(y)$ respectively. The cdf of $X+Y$ is called the *convolution* of $F_X(x)$ and $F_Y(y)$ and is defined as

$$F_{X+Y}(a) = P(X+Y \le a) = \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\, dy$$

Definition 4.17 (Differentiation of the Convolution)**.** The pdf of $X+Y$ is obtained by differentiating the *convolution* of $F_X(x)$ and $F_Y(y)$:

$$f_{X+Y}(a) = \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\, dx = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\, dy$$

Example 4.20. Let $X \sim \exp\left(\frac{1}{2}\right), Y \sim \exp\left(\frac{1}{4}\right)$ be independent. We want to find $f_{X+Y}(z) = f_Z(z)$ where $Z = X+Y$.

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy$$
$$f_X(z-y) = \begin{cases} \frac{1}{2} e^{-\frac{z-y}{2}} & 0 \le z-y \le \infty \\ 0 & \text{otherwise} \end{cases} = \begin{cases} \frac{1}{2} e^{-\frac{z-y}{2}} & y \ge z \\ 0 & \text{otherwise} \end{cases}$$
$$f_Y(y) = \begin{cases} \frac{1}{4} e^{-\frac{1}{4}(y)} & 0 \le y \le \infty \\ 0 & \text{otherwise} \end{cases}$$

We have the same bounds for y over all possible valuez of z, so we need to compute only 1 integral.

5 Multivariate Random Variables

5.1 Lecture 14: Law of Large Numbers, Central Limit Theorem, Convergence in Distribution

Suppose we want to solve the following problem:

1. Let $X_1, \ldots, X_n$ be n independent random variables.
2. Let $u_1(\cdot), \ldots, u_n(\cdot)$ be n functions.
3. Let $Y = u_1(X_1) + u_2(X_2) + \cdots + u_n(X_n) = \sum_{i=1}^n u_i(X_i)$

What exactly is $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$? Let $u_1(X_1) = a_1X_1, \ldots, u_n(X_n) = a_nX_n$, $a_i \in \mathbb{R} \forall i$.

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}\left[\sum_{i=1}^n a_iX_i\right] = \mathbb{E}[a_1X_1] + \mathbb{E}[a_2X_2] + \cdots + \mathbb{E}[a_nX_n] \\ &= a_1\mathbb{E}[X_1] + a_2\mathbb{E}[X_2] + \cdots + a_n\mathbb{E}[X_n] \\ &= a_1\mu_{X_1} + \cdots + a_n\mu_{X_n} = \sum_{i=1}^n a_i\mu_{X_i}\end{aligned}$$

Note $\forall i = 1, \ldots, n$,

$$\begin{aligned}\mathbb{E}\left[(a_iX_I - a_i\mu_{X_i})^2\right] &= \mathbb{E}\left[(a_iX_i)^2 - 2a_i^2X_i\mu_{X_i} + a_i^2\mu_{X_i}^2\right] \\ &= a_i^2\mathbb{E}\left[X_i^2\right] - 2a_i^2\mu_{X_i}^2 + a_i^2\mu_{X_i}^2 = a_i^2\left(\mathbb{E}\left[X_i^2\right] - \mu_{X_I}^2\right) \\ &= a_i^2\sigma_{X_i}^2 \qquad \text{(Result 1)}\end{aligned}$$

Also, note that $X_1, \ldots, X_n$ are independent. Let us consider for now $a_1 = a_2 = \cdots = a_n = 1$, such that $Y = \sum_{i=1}^n X_i$:

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{i=1}^n \mu_{X_i} \\ \sigma_Y^2 &= \mathbb{E}\left[\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_{X_i}\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum X_i\right)^2 - 2\left(\sum X_i\right)\left(\sum \mu_{X_i}\right) + \left(\sum \mu_{X_i}\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum X_i\right)^2\right] - 2\left(\sum \mu_{X_i}\right)\mathbb{E}\left[\sum X_i\right] + \left(\sum \mu_{X_i}\right)^2 \\ &= \mathbb{E}\left[\left(\sum X_i\right)^2\right] - \left(\sum \mu_{X_i}\right)^2\end{aligned}$$

Let $i = 1, 2$, i.e., $n = 2$, then

$$\begin{aligned}\sigma_Y^2 &= \mathbb{E}\left[(X_1 + X_2)^2 - (\mu_{X_1} + \mu_{X_2})^2\right] \\ &= \mathbb{E}\left[X_1^2 + 2X_1X_2 + X_2^2 - \mu_{X_1}^2 - 2\mu_{X_1}\mu_{X_2} - \mu_{X_2}^2\right] \\ &= \mathbb{E}\left[X_1^2\right] + \mathbb{E}\left[X_2^2\right] - \mu_{X_1}^2 - \mu_{X_2}^2 + 2\mathbb{E}[X_1]\mathbb{E}[X_2] - 2\mu_{X_1}\mu_{X_2} \\ &= \mathbb{E}\left[X_1^2\right] - \mu_{X_1}^2 + \mathbb{E}\left[X_2^2\right] - \mu_{X_2}^2 \\ &= \sigma_{X_1}^2 + \sigma_{X_2}^2\end{aligned}$$

This result holds for any $n > 2$, then,

$$Y = \sum_{i=1}^{n} X_i \implies \sigma_Y^2 = \sum_{i=1}^{n} \sigma_{X_i}^2 \qquad \text{(Result 2)}$$

Combining *Result 1* and *Result 2*, we get the fundamental result that

$$Y = \sum_{i=1}^{n} a_i X_I \implies \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_{X_i}^2$$

Theorem 5.1. *Let $X_1, \ldots, X_n$ be n independent random variables. Let $a_1, \ldots, a_n \in \mathbb{R}$. Let Y be the random variable*

$$Y = \sum_{i=1}^{n} a_i X_i = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

Then,

$$\mu_Y = \mathbb{E}[Y] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i] = a_1 \mathbb{E}[X_1] + a_2 \mathbb{E}[X_2] + \cdots + a_n \mathbb{E}[X_n]$$

$$\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_{X_i}^2 = a_1^2 \sigma_{X_1}^2 + a_2^2 \sigma {X_2}^2 + \cdots + a_n^2 \sigma {X_n}^2$$

Example 5.1. Let X_1, X_2, X_2 be independent binomial random variables such that

$$X_1 \sim \text{bin}(n_1 = 4, p_1 = 1/2)$$
$$X_2 \sim \text{bin}(n_2 = 6, p_2 = 1/3)$$
$$X_3 \sim \text{bin}(n_3 = 12, p_3 = 1/6)$$

Let $Y = X_1 + 2X_2 + 3X_3$. Then,

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3] \\
&= n_1 p_1 + 2 n_2 p_2 + 3 n_3 p_3 \\
&= 12 \\
\sigma_Y^2 &= \sigma_{X_1}^2 + 4\sigma_{X_2}^2 + 9\sigma_{X_3}^2 \\
&= n_1 p_1 (1 - p_1) + 4 n_2 p_2 (1 - p_2) + 9 n_3 p_3 (1 - p_3) \\
&= 64/3
\end{aligned}$$

Example 5.2. Let $a_1 = a_2 = \cdots = a_n = \frac{1}{n}$ and $X_1, \ldots, X_n$ be n independent random variables. Then,

$$\tilde{X} = \sum_{i=1}^{n} a_i X_i = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Let us also consider the case where $X_1, \ldots X_n$ are identically distributed with mean μ and variance σ^2. Then, by Theorem 5.1, we have

$$\mathbb{E}\left[\tilde{X}\right] = \frac{1}{n} \sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

$$\sigma_{\tilde{X}}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Theorem 5.2 (Weak Law of Large Numbers)**.** *Let $X_1, \ldots, X_n$ be n independent random variables with mean μ and variance σ^2. Let*

$$\tilde{X} = \frac{X_1 + \cdots + X_n}{n}$$

Then,

$$\mathbb{E}\left[\tilde{X}\right] = \mu \qquad Var\left(\tilde{X}\right) = \frac{\sigma^2}{n}$$

Remark. The importance of the Law of Large Numbers is given by its implication that the result of one random experiment is unforseeable, but the average result of a large number of iterations of the same random experiment is predictable.

$$\lim_{n\to\infty} \mathbb{E}\left[\tilde{X}\right] = \mu, \qquad \lim_{n\to\infty} \mathrm{Var}\left(\tilde{X}\right) = 0$$

Example 5.3. Let $Y = a_1X_1 + \cdots + a_nX_n$ where $X_1, \ldots, X_n$ are independent random variables and $a_1, \ldots, a_n \in \mathbb{R}$. We want to find $M_Y(t)$, the moment generating function of Y.

$$\begin{aligned} M_Y(t) &= \mathbb{E}[\exp(ty)] = \mathbb{E}[\exp(t(a_1X_1 + \cdots + a_nX_n))] \\ &= \mathbb{E}[\exp(ta_1X_1)\exp(ta_2X_2)\cdots\exp(ta_nX_n)] \\ &= \mathbb{E}[\exp(ta_1X_1)]\,\mathbb{E}[\exp(ta_2X_2)]\cdots\mathbb{E}[\exp(ta_nX_n)] \\ &= M_{X_1}(a_1t)\,M_{X_2}(a_2t)\cdots M_{X_n}(a_nt) \\ &= \prod_{i=1}^{n} M_{X_i}(a_it) \end{aligned}$$

We can express this succinctly as

$$\begin{aligned} M_Y(t) = \mathbb{E}[\exp(ty)] &= \mathbb{E}\left[\exp\left(t\left(\sum_{i=1}^{n} a_iX_i\right)\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n}\exp(ta_iX_i)\right] \\ &= \prod_{i=1}^{n}\mathbb{E}[\exp(ta_iX_i)] = \prod_{i=1}^{n} M_{X_i}(a_it) \end{aligned}$$

Theorem 5.3. *Let $X_1, \ldots, X_n$ be n independent random variables. Let $a_1, \ldots, a_n \in \mathbb{R}$. Let Y be the random variable $Y = \sum_{i=1}^{n} a_iX_i$. Then,*

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_it)$$

Example 5.4. Let $X_1, \ldots, X_n$ be n independent exponential random variables with mean θ. Let $Y = \sum_{i=1}^{n} X_i$. We know that $M_{X_i} = \frac{1}{(1-\theta t)}$, then,

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \frac{1}{(1-\theta t)^n}$$

the moment generating function of a Gamma random variable with parameters n, θ.

Example 5.5. Let $X_1, \ldots, X_n$ be n independent and identically distributed normal random variables:

$$X_i \sim N\left(\mu, \sigma^2\right) \quad \forall i \in \{1, \ldots, n\}$$

Let $\tilde{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \implies \mathbb{E}\left[\tilde{X}\right] = \mu$ and $\text{Var}\left(\tilde{X}\right) = \frac{\sigma^2}{n}$. We want to understand $M_{\tilde{X}}(t)$, the moment generating function of $\tilde{X}$. Now,

$$M_{\tilde{X}}(t) = \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{n}\right)$$

and

$$M_{X_i}(t) = \exp\left[\mu t + \frac{t^2}{2}\sigma^2\right] \implies M_{X_i}\left(\frac{t}{n}\right) = \exp\left[\mu\frac{t}{n} + \frac{t^2}{2n^2}\sigma^2\right]$$

Then,

$$\begin{aligned} M_{\tilde{X}}(t) &= \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{n}\right) = \prod_{i=1}^{n} \exp\left[\mu\frac{t}{n} + \frac{t^2}{2n^2}\sigma^2\right] = \left[\exp\left[\mu\frac{t}{n} + \frac{t^2}{2n^2}\sigma^2\right]\right]^n \\ &= \exp\left[n\left[\mu\frac{t}{n} + \frac{t^2}{2n^2}\sigma^2\right]\right] = \exp\left(\mu t + \frac{t^2}{2}\frac{\sigma^2}{n}\right) \end{aligned}$$

The moment generating function of a normal random variable with parameters $\left(\mu, \sigma^2/n\right)$, then,

$$\tilde{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Then,

$$W = \frac{\tilde{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

If $X_1, \ldots, X_n$ are independent and identically distributed with mean μ and variance σ^2, even if they are not normal random variables, it still holds that

$$\lim_{n\to\infty} W = \lim_{n\to\infty} \frac{\tilde{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim N(0, 1)$$

where $\xrightarrow{d}$ indicates convergence in distribution.

Theorem 5.4 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be independent random variables with mean μ and variance σ^2. Let W be the random variable*

$$W = \frac{\tilde{X} - \mu}{\sigma/\sqrt{n}} = \frac{\left(\sum_{i=1}^{n} X_i\right) - n\mu}{\sigma\sqrt{n}}$$

then W converges to a standard normal random variable as $n \longrightarrow \infty$.

Example 5.6. Let $X_1, \ldots X_{500}$ be the change in value of 500 sticks in a time span of 1 year. Assume that $X_1, \ldots, X_{500}$ have the same mean $\mu = 0$, and $\sigma^2 = 64$, a very unrealistic assumption. Also assume that $X_1, \ldots, X_{500}$ are independent, also an unrealistic assumption. We ask, how much money can we

make if we purchase onf each stock one year from now? We cannot answer this question because it would mean that we are able to predict the outcome of each one of these 500 random experiments. However, we can ask the probability that we will make a profit of at least 2?

$$P\left(\tilde{X} \geq 2\right) = P\left(\frac{\tilde{x}-0}{8/500} \geq \frac{2-0}{8/500}\right) = 1 - P\left(\frac{\tilde{x}-0}{8/500} < \frac{1000}{8}\right)$$
$$= 1 - \phi\left(125\right) \approx 1 - 1 \approx 0$$

because, by Central Limit Theorem, we have that $\frac{\tilde{x}-0}{8/500} \sim N\left(0,1\right)$.

Remark. Remember when we showed that a binomial random variable with parameters $\left(n, \frac{\lambda}{n}\right)$ converges to a Poisson with parameter λ? That was an example of convergence in distribution.

$$X \sim \text{bin}\left(n, \frac{\lambda}{n}\right) \implies f_X\left(k\right) = P\left(X = k\right) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k}$$
$$Y \sim \text{Poisson}\left(\lambda\right) \implies f_Y\left(k\right) = P\left(Y = k\right) = \frac{\lambda^k e^{-\lambda}}{k!}$$

and we showed that

$$\lim_{n\to\infty} f_X\left(k\right) = f_Y\left(k\right)$$

Example 5.7. Let us consider the sequnce of continuous, independent random variables $X_1, \ldots, X_n$. Let

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

Let's assume that the distribution of $X_1, \ldots, X_n$ is such that

$$F_Y\left(x\right) = \begin{cases} 1 - \left(1 - \frac{1}{n}\right)^{nx} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Show that Y converges in distribution to an exponential with parameter $\lambda = 1$. Remember that $W \sim \exp\left(1\right) \implies F_W\left(x\right) = 1 - e^{-x}$. Then,

$$\lim_{n\to\infty} F_Y\left(x\right) = \lim_{n\to\infty}\left(1 - \left(1 - \frac{1}{n}\right)^{nx}\right) = 1 - \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^{nx}$$
$$= 1 - e^{-x} = F_W\left(x\right)$$

so $y \xrightarrow{d} W \sim \exp\left(1\right)$, i.e., Y converges in distribution to an exponential random variable with parameter $\lambda = 1$.

Definition 5.1 (Convergence in Distribution)**.** A sequence of random variables $X_1, \ldots, X_n$ converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_{X_n}\left(x\right) = F_X\left(x\right)$$

where $F_{X_n}\left(x\right)$ is the cdf of the sequence $X_1, \ldots, X_n \;\; forallx$ at which the cdf of X, $F_X\left(x\right)$ is continuous.

In some cases, if $X_1, \ldots, X_n$ are discrete random variables, there might be an issue with the Central Limit Theorem. let $X_1 \ldots, X_n$ be independent and identically distributed Bernoulli random variables with parameter P. Then,

$$Y = \sum_{i=1}^{n} X_i \sim \text{bin}(n, p) \qquad \mathbb{E}[Y] = np \qquad \text{Var}(Y) = np(1-p)$$

By Central Limit Theorem,

$$W = \frac{Y - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z \sim N(0, 1)$$

Let $k \in \{1, 2, \ldots, n\}$. Then, by Central Limit Theorem,

$$P(Y = k) = P\left(W = \frac{k - np}{\sqrt{np(1-p)}}\right)$$

We is continuous, so the RHS is always 0. But, the LHS is not. This contradiction arises because we are approximatting a discrete random variable with a continuous one. We can solve this issue using a "half-unit correction". Because k is an integer

$$P(Y = k) = P(k - \varepsilon < Y < k + \varepsilon), \quad \varepsilon \in \mathbb{R}$$

Thenm

$$Y = k - \varepsilon \implies W = \frac{k - \varepsilon - np}{\sqrt{np(1-p)}}$$
$$Y = k + \varepsilon \implies W = \frac{k + \varepsilon - np}{\sqrt{np(1-p)}}$$

By Central Limit Theorem,

$$P(Y = k) = P\left(\frac{k - \varepsilon - np}{\sqrt{np(1-p)}} < W < \frac{k + \varepsilon - np}{\sqrt{np(1-p)}}\right)$$

This method's name comes from the fact that when it was first proposed, $\varepsilon = 0.5$.