Basics of Probability Theory

Yanze Song

December 10, 2023

1 Probability Basics

1.1 Terminologies

- Ramdom Experiment: Anything whose outcome is unknown until occurred. Rolling a 6-sided dice is an example: we know the outcome (defined as the point faced-up) will be in $\{1, 2, ..., 6\}$, but can not tell which one exactly will be faced up until the die is rolled and stops.
- Sample Point: A single outcome of a random experiment. Exactly one sample point will occur on any trail of a ramdom experiment.
- Sample Space (Ω): All sample points of a random experiment. There are discrete and continuous sample space. We will focus on the discrete one first, defined such that the sample points are countable.
- Event: A subset of the sample space, or a set of sample points. It could be empty.

When describing a random experiment, we need to clearly define what the sample points are. For example, when rolling a die, we could define the sample points as how long the die is spinning, instead of the point faced up when it stops.

1.2 Set Theory

Since events are defined as sets, the following formulas in the set theory can be found useful:

- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$.
- $\overline{A \cup B} = \overline{A} \cap \overline{B}$.

1.3 Probability Axioms

Definition 1.3.1 (*Probability Axioms*)**.** *W.r.t. a random experiment, for any event A, we assign a number $P(A)$ called the probability of A, so that the followings hold:*

- *Axiom 1: $0 \leq P(A) \leq 1$.*
- *Axiom 2: $P(\Omega) = 1$.*
- *Axiom 3: Given that $A_1, A_2, ...$ are pairwise disjoint, $P(\bigcup\limits_{i=1}^{\infty} A_i) = \sum\limits_{i=1}^{\infty} P(A_i)$.*

Rigorously, $P : \mathcal{P}(\Omega) \to [0, 1]$ is called a probability measure.

Lemma 1.3.1 (*Immediate Results*)**.** *Here are some immediate results from the axioms:*

- *$P(\overline{A}) = 1 - P(A)$.*
- *$P(\emptyset) = 0$.*
- *Given that $A_1, A_2, ...A_n$ are pairwise disjoint, $P(\bigcup\limits_{i=1}^{n} A_i) = \sum\limits_{i=1}^{n} P(A_i)$.*

1.4 Sample-Point Method

The goal of defining probability is to quantify the likelihood of the events, and here we are! The following is called the *Sample-Point Method*.

- Define a *probability model*: Define what the sample points are and their associated probabilities so that the first two axioms hold.
- Express the event A as a set of sample points.
- Calculate $P(A)$ as $P(A) = \sum\limits_{i:i \in A} P(i)$, where i is a sample point.

The above three axioms hold if we follow this method. We don't randomly assign probability to the sample points. One interpretation of probability is that in the long run, the *relative frequency* of an event (defined as $\frac{n_E}{N} = \frac{\text{\# of times } E \text{ is observed}}{\text{total \# of trials}}$) should approach to this value.

Theorem 1.4.1 (*Equiprobable Sample Points*)**.** *Given Ω with $|\Omega| = N$, assume all sample points are equiprobable, i.e. they have the same probability, then the probability of each sample point is*

$$P(E_i) = \frac{1}{N},\ \forall i.$$

More generally, for an event A of n_A equiprobable sample points,

$$P(A) = \frac{n_A}{N}.$$

Based on this assumption, finding the probability of an event becomes finding the number of sample points in the sample space and the event. This is where counting methods come in.

1.5 Counting Methods

Theorem 1.5.1 (*Multiplication Rule*)**.** *Consider r experiments where experiment i has n_i outcomes, then there are $n_1 n_2 ... n_r$ outcomes for the r experiments.*

Theorem 1.5.2 (*Permutations*)**.** *The number of permutations (order matters, without replacement) of r out of n distinct objects is*

$$P_r^n = \frac{n!}{(n-r)!}.$$

Theorem 1.5.3 (*Combinations*)**.** *The number of combinations (order doesn't matter, without replacement) of r out of n distinct objects is*

$$\binom{n}{r} = \frac{P_r^n}{r!} = \frac{n!}{(n-r)! \cdot r!}$$

Theorem 1.5.4 (*Multinomial Combinations*)**.** *The number of ways to divide n distinct objects into k distinct groups of size $n_1, n_2, ..., n_k$ such that $\sum_{k=1}^{n} n_k = n$, where the order within each group doesn't matter is*

$$\binom{n}{n_1\ n_2\ ...\ n_k} = \frac{n!}{n_1! \cdot n_2! \cdot ... \cdot n_k!}.$$

Equivalently, the following also works

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}...\binom{n-n_1-...-n_{k-1}}{n_k}.$$

1.6 Conditional Probability

Definition 1.6.1 (*Conditional Probability*)**.** *The conditional probability of event A, given that event B has occurred, is defined as*

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

provided that $P(B) > 0$.

$P(A)$ can be written as $P(A|\Omega) = \frac{P(A\cap\Omega)}{P(\Omega)} = \frac{P(A)}{1} = P(A)$, measuring the likelihood of A given that something will occur. Regarding $P(A|B)$, we can consider that the sample space shrinks from Ω to B, measuring the likelihood of A, given that B has occurred. Based on this understanding, we know that $A|B$ and $\overline{A}|B$ form a probability distribution as well, i.e.,

$$P(A|B) + P(\overline{A}|B) = 1.$$

Definition 1.6.2 (*Partition*)**.** *Let $\{E_i\}_{i=1}^n$ be a set of events, then $\{E_i\}_{i=1}^n$ is defined to be a partition of the sample space Ω if the followings hold:*

- $\bigcup\limits_{i=1}^{n} E_i = \Omega$
- $\forall i \neq j : E_i \cap E_j = \emptyset$

Theorem 1.6.1 (*Total Probability*)**.** *Let $\{E_i\}_{i=1}^n$ be a partition of the sample space Ω, then for any event A, we have*

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

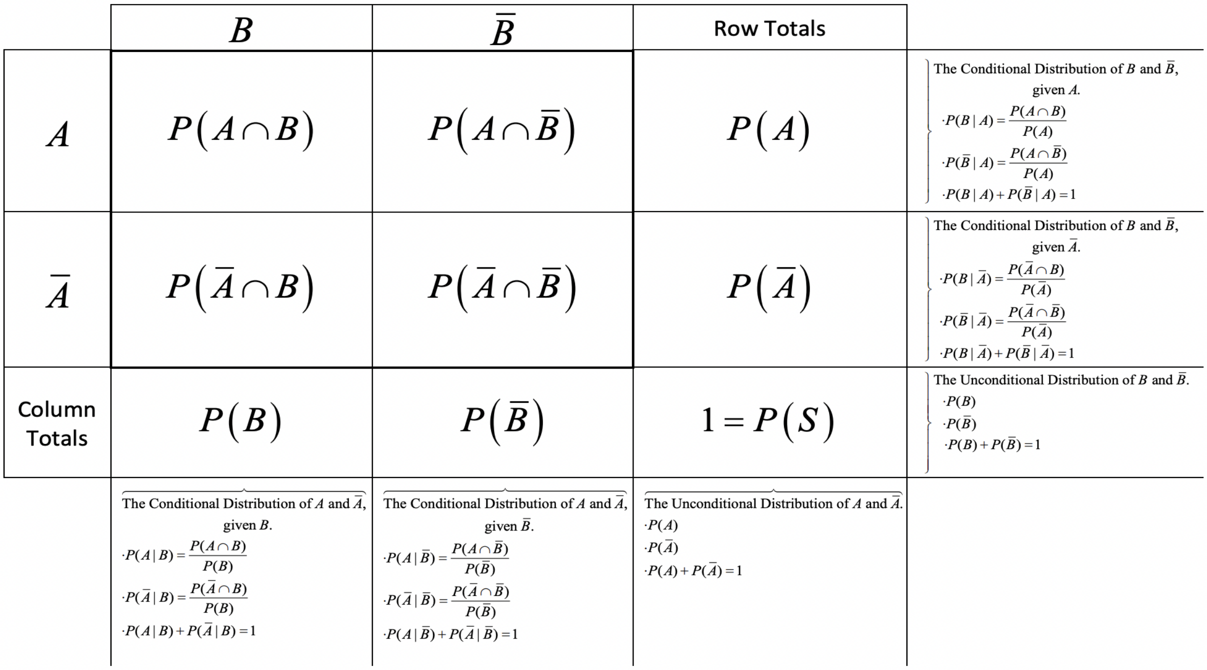
More specifically, $P(A) = P(A \cap B) + P(A \cap \overline{B})$.

Proof.

$$\begin{aligned} P(A) = P(A \cap \Omega) &= P(A \cap \bigcup_{i=1}^{n} E_i) \\ &= P(\bigcup_{i=1}^{n} A \cap E_i) \\ &= \sum_{i=1}^{n} P(A \cap E_i), \text{ by Axiom 3.} \end{aligned}$$

□

Definition 1.6.3 (*Contingency Table*)**.** *Below is the contingency table of events A and B:*



	B	$\overline{B}$	Row Totals	
A	$P\left(A\cap B\right)$	$P\left(A\cap \overline{B}\right)$	$P\left(A\right)$	The Conditional Distribution of B and $\overline{B}$, given A. $\cdot P(B\mid A)=\frac{P(A\cap B)}{P(A)}$ $\cdot P(\overline{B}\mid A)=\frac{P(A\cap \overline{B})}{P(A)}$ $\cdot P(B\mid A)+P(\overline{B}\mid A)=1$
$\overline{A}$	$P\left(\overline{A}\cap B\right)$	$P\left(\overline{A}\cap \overline{B}\right)$	$P\left(\overline{A}\right)$	The Conditional Distribution of B and $\overline{B}$, given $\overline{A}$. $\cdot P(B\mid \overline{A})=\frac{P(\overline{A}\cap B)}{P(\overline{A})}$ $\cdot P(\overline{B}\mid \overline{A})=\frac{P(\overline{A}\cap \overline{B})}{P(\overline{A})}$ $\cdot P(B\mid \overline{A})+P(\overline{B}\mid \overline{A})=1$
Column Totals	$P\left(B\right)$	$P\left(\overline{B}\right)$	$1=P\left(S\right)$	The Unconditional Distribution of B and $\overline{B}$. $\cdot P(B)$ $\cdot P(\overline{B})$ $\cdot P(B)+P(\overline{B})=1$
	The Conditional Distribution of A and $\overline{A}$, given B. $\cdot P(A\mid B)=\frac{P(A\cap B)}{P(B)}$ $\cdot P(\overline{A}\mid B)=\frac{P(\overline{A}\cap B)}{P(B)}$ $\cdot P(A\mid B)+P(\overline{A}\mid B)=1$	The Conditional Distribution of A and $\overline{A}$, given $\overline{B}$. $\cdot P(A\mid \overline{B})=\frac{P(A\cap \overline{B})}{P(\overline{B})}$ $\cdot P(\overline{A}\mid \overline{B})=\frac{P(\overline{A}\cap \overline{B})}{P(\overline{B})}$ $\cdot P(A\mid \overline{B})+P(\overline{A}\mid \overline{B})=1$	The Unconditional Distribution of A and $\overline{A}$. $\cdot P(A)$ $\cdot P(\overline{A})$ $\cdot P(A)+P(\overline{A})=1$	

where $P(A\cap B)$ is called the joint probability, and $P(A)$ is called the marginal probability. In this case, the 4 joint probabilities form a probability distribution, i.e, $P(A\cap B)+P(A\cap \overline{B})+P(\overline{A}\cap B)+P(\overline{A}\cap \overline{B})=1$, and 2 marginal probabilities form one as well, i.e, $P(A)+P(\overline{A})=P(B)+P(\overline{B})=1$.

1.7 Independence

What does it mean by $P(A|B) = P(A)$? The occurrence of B doesn't affect the probability of the occurrence of A, and we call this relation *independence*.

Definition 1.7.1 (*Independence*)**.** *Two events A and B are said to be independent if any of the followings holds:*

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \cap B) = P(A) \cdot P(B)$

They are called dependent otherwise.

The above three statements are equivalent. To see this,

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

Lemma 1.7.1. *A and B are independent if and only if*

$$\exists E_1 \in \{A, \overline{A}\}, \exists E_2 \in \{B, \overline{B}\} : \{E_1, E_2\} \text{ are independent,}$$

where $\{E_1, E_2\}$ means that the order doesn't matter.

Proof. To simplify, we will only prove B and $\overline{A}$ are independent $\Leftrightarrow$ A and B are independent as an exercise:

$$\begin{aligned} P(B|\overline{A}) = P(B) &\Leftrightarrow \frac{P(B \cap \overline{A})}{P(\overline{A})} = P(B) \\ &\Leftrightarrow \frac{P(B) - P(B \cap A)}{1 - P(A)} = P(B) \\ &\Leftrightarrow P(A \cap B) = P(A) \cdot P(B). \end{aligned}$$

□

Theorem 1.7.1 (*Multiplication Rule*)**.** *The probability of the intersection of two events A and B is*

$$P(A \cap B) = P(A|B) \cdot P(B).$$

Note if A and B are independent, then $P(A \cap B) = P(A) \cdot P(B)$. More generally, for three events A, B and C, we have

$$P(A \cap B \cap C) = P(A|B \cap C) \cdot P(B \cap C) = P(A|B \cap C) \cdot P(B|C) \cdot P(C).$$

Theorem 1.7.2 (*Total Probability*)**.** *Let $\{E_i\}_{i=1}^n$ be a partition of the sample space S, then for any event A, we have*

$$P(A) = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(A|E_i) \cdot P(E_i).$$

More specifically, $P(A) = P(A \cap B) + P(A \cap \overline{B}) = P(A|B) \cdot P(B) + P(A|\overline{B}) \cdot P(\overline{B})$.

Remark 1.7.1. The grouping of the intersections matters, that is,

$$\begin{aligned} P(A \cap B) &= P(A|B) \cdot P(B) \\ &= P(B|A) \cdot P(A). \end{aligned}$$

The above are all the same, but some are known and some are unknown. We should use the ordering where all the terms are known to us. Same if there're 3 events:

$$\begin{aligned} P(A \cap B \cap C) &= P(A|B \cap C) \cdot P(B|C) \cdot P(C) \\ &= P(C|A \cap B) \cdot P(A|B) \cdot P(B) \\ &= ... \end{aligned}$$

Theorem 1.7.3 (*Inclusion-Exclusion Rule*)**.** *The probability of the union of two events A and B is*

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

Note if A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$ by Axiom 3. For three events A, B and C, we have

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

More generally, for an arbitrary n events, we have

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i) - \sum_{i \neq j} P(E_i \cap E_j) + \sum_{i \neq j \neq k} P(E_i \cap E_j \cap E_k) - ...$$

2 Discrete Random Variables

2.1 Discrete Random Variables Basics

Definition 2.1.1 (*Random Variable*)**.** *A random variable (RV) is a function $X : \Omega \to \mathbb{R}$ that assigns a real number to each sample point, i.e. $\forall \omega \in \Omega : X(\omega) \in \mathbb{R}$. X is discrete if range(X) is countable. Here's a common notation:*

$$X = k \coloneqq \{\omega \in \Omega \,|\, X(\omega) = k\}.$$

Remark 2.1.1. W.r.t. a random experiment, a random variable *numerically* categorizes the sample points based on its definition, after the sample point is clearly defined. The definition of the sample point comes first, and then the random variable.

Consider flipping a coin twice, and define the sample point as the sides faced up. Therefore,

$$\Omega = \{HH, HT, TH, TT\}.$$

Define a random variable X as the number of heads, then

$$\Omega = \{\underbrace{HH}_{X=2}, \underbrace{HT, TH}_{X=1}, \underbrace{TT}_{X=0}\}.$$

We switch from words to random variables to describe events in a more mathematically rigorous manner. However, they both represent of a set of sample points in essence.

Definition 2.1.2 (*Probability Mass Function*)**.** *Let X be a discrete random variable. The probability distribution of X is the set of probabilities $\{P(X = x)\}_{all\ x}$. $P_X(\cdot)$ is called the probability mass function (pmf) of X.*

Remark 2.1.2. A few notations:

- $P_X(x) \coloneqq P(X = x)$
- $P(X = x, Y = y) \coloneqq P(X = x \cap Y = y)$

Lemma 2.1.1 (*Immediate Results*)**.** *For any pmf $P_X(\cdot)$, the followings must be true:*

- $\forall x : P_X(x) \in [0, 1]$
- $P(\bigcup\limits_{all\ x} X = x) = P(\Omega) = 1$
- $P(X = x_1 \cup X = x_2 \cup ... \cup X = x_n) = \sum\limits_{i=1}^{n} P(X = x_i)$

2.2 Expectations

Definition 2.2.1 (*Raw Moments*)**.** *Let X be a discrete RV, then the k^{th} raw moment, or the moments about the origin of X is defined as*

$$\mathbb{E}[X^k] \; = \sum_{all\ x} x^k \cdot P(X = x).$$

Definition 2.2.2 (*Expectations*)**.** *The expected value or mean of X is defined as the first raw moment of X:*

$$\mu_X = \mathbb{E}[X] = \sum_{all\ x} x \cdot P(X = x).$$

It measures the average of the distribution weighted by their probabilities, due to their different frequencies of occurrences.

Theorem 2.2.1 (*Composition on Expectations*)**.** *Let g be a real-valued function of X, then*

$$\mathbb{E}[g(X)] = \sum_{all\ x} g(x) \cdot P(X = x).$$

Theorem 2.2.2 (*Properties of Expectations*)**.** *Let $\{g_i\}_{i=1}^n$ be real-valued functions of X, and $\{c_i\}_{i=1}^n$ be constants. Then, we have*

- $\mathbb{E}[c] = c$
- $\mathbb{E}[cg(X)] = cE[g(X)]$
- $\mathbb{E}[g_1(X) + ... + g_n(X)] = \mathbb{E}[g_1(X)] + ... + \mathbb{E}[g_n(X)]$

Proof. We'll equivalently prove the following:

$$\mathbb{E}[\sum_{i=1}^n c_i \cdot g_i(X)] = \sum_{i=1}^n c_i \cdot \mathbb{E}[g_i(X)].$$

$$\begin{aligned} \mathbb{E}[\sum_{i=1}^n c_i \cdot g_i(X)] &= \sum_{\text{all x}} (\sum_{i=1}^n c_i \cdot g_i(x)) \cdot P(X = x) \\ &= \sum_{i=1}^n c_i \cdot (\sum_{\text{all x}} g_i(x) \cdot P(X = x)) \\ &= \sum_{i=1}^n c_i \cdot \mathbb{E}[g_i(X)] \end{aligned}$$

□

Theorem 2.2.3 (*Scaling & Shifting on Expectations*)**.** *Let a, b be constants, we have the following:*

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

Proof. We'll use the definition of expectations for the proof:

$$\begin{aligned}
\mathbb{E}[aX+b] &= \sum_{\text{all x}} (ax+b)\cdot P(X=x) \\
&= \sum_{\text{all x}} ax\cdot P(X=x) + b\cdot P(X=x) \\
&= \sum_{\text{all x}} ax\cdot P(X=x) + \sum_{\text{all x}} b\cdot P(X=x) \\
&= a\underbrace{\sum_{\text{all x}} x\cdot P(X=x)}_{\mathbb{E}[X]} + b\underbrace{\sum_{\text{all x}} P(X=x)}_{1} \\
&= a\mathbb{E}[X]+b
\end{aligned}$$

□

Remark 2.2.1. What does $\mathbb{E}[X]=a$ mean? Suppose we run a random experiment a lot of times, and for each outcome we calculate $X(\omega)$. In the end, after adding up all $X(\omega)$ and taking the average, we get a.

2.3 Variance

Definition 2.3.1 (*Central Moments*)**.** *The k^{th} central moment, or the moment about the mean of X is defined as*

$$\mathbb{E}[(X - \mu_X)^k] = \sum_{all\ x} (y - \mu_X)^k \cdot P(X = x).$$

Remark 2.3.1 (*Zero First Central Moment*)**.** The first central moment of X is zero, i.e.,

$$\mathbb{E}[X - \mu_X] = \mathbb{E}[X] - \mathbb{E}[\mu_X] = \mu_X - \mu_X = 0.$$

That is, throughout the distribution, there're some points above the mean, and some are below the mean, but the weighted average viability around the mean is zero.

Definition 2.3.2 (*Variance*)**.** *The variance of X is defined as the second central moment of X:*

$$\sigma_X^2 = Var[X] = \mathbb{E}[(X - \mu_X)^2] = \sum_{all\ x} (x - \mu_X)^2 \cdot P(X = x).$$

It measures the Euclidean distance between all points and the mean weighted by their probabilities, and how spread out around the mean the whole distribution is.

Theorem 2.3.1 (*Composition on Variance*)**.** *Let g be a real-valued function of X, then*

$$Var[g(X)] = \mathbb{E}[(g(X) - \mathbb{E}[g(X)])^2] = \sum_{all\ x} (g(x) - \mathbb{E}[g(X)])^2 \cdot P(X = x).$$

Theorem 2.3.2 (*Variance Using Expectations*)**.** *We can calculate the variance of X by*

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

More generally, $Var[g(X)] = \mathbb{E}[g(X)^2] - \mathbb{E}[g(X)]^2$.

Proof.

$$\begin{aligned} Var[X] &:= \mathbb{E}[(X - \mu_X)^2] \\ &= \mathbb{E}[X^2 - 2\mu_X X + \mu_X^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[\mu_X X] + E[\mu_X^2] \\ &= \mathbb{E}[X^2] - \mu_X^2 \end{aligned}$$

□

Theorem 2.3.3 (*Shifting & Scaling on Variance*)**.** *Let a, b be constants, we have the following:*

$$Var[aX + b] = a^2 Var[X].$$

In particular, $Var[aX] = a^2 Var[X]$, and $Var[b] = 0$.

Proof.

$$\begin{aligned}
Var[aX+b] &= \sum_{\text{all x}} (ax+b-\mathbb{E}[aX+b])^2 \cdot P(X=x) \\
&= \sum_{\text{all x}} (ax+b-a\mathbb{E}[X]-b)^2 \cdot P(X=x) \\
&= \sum_{\text{all x}} (ax-a\mathbb{E}[X])^2 \cdot P(X=x) \\
&= a^2 \sum_{\text{all x}} (x-\mathbb{E}[X])^2 \cdot P(X=x) \\
&= a^2 Var[X]
\end{aligned}$$

□

Remark 2.3.2. It makes sense that the shifting (b) doesn't affect the variability, because the whole distribution shifts together and the shape is maintained.

Definition 2.3.3 (*Standard Deviation*)**.** *The standard deviation of X is defined as*

$$\sigma_X = \sqrt{Var[X]}.$$

Remark 2.3.3. What if $\mathbb{E}[X^2] = \mathbb{E}[X]^2$? Then $Var(X) = 0$, meaning that there's no variability around the mean, i.e., all points concentrate on the mean point. More rigorously, if we only consider x s.t. $P(X=x) > 0$, then

$$\begin{aligned}
&Var[X] = \sum_{\text{all x}} (x-\mu_X)^2 \cdot P(X=x) = 0 \\
&\Rightarrow (x-\mu_X)^2 = 0, \text{ for all x, because } P(X=x) > 0 \\
&\Rightarrow x = \mu_X, \text{ for all x}
\end{aligned}$$

In this case, X is a constant, or a constant function, since range$(X) = \{\mu_X\}$. It's sometimes considered not an RV (no randomness here: $X(\omega) = \mu_X, \forall \omega \in \Omega$), or a degenerate RV.

2.4 I.I.D. Random Variables

Definition 2.4.1 (*Independence of Random Variables*). *Let X and Y be random variables, then we say X and Y are independent if one of the followings holds:*

- $\forall x, \forall y : P(X = x \,|\, Y = y) = P(X = x)$
- $\forall x, \forall y : P(Y = y \,|\, X = x) = P(Y = y)$
- $\forall x, \forall y : P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$

Remark 2.4.1. The above three statements are equivalent. It mimics the definition of independent events.

Definition 2.4.2 (*Identical Distribution*). *X and Y are called identically distributed if*

$$P_X = P_Y.$$

Remark 2.4.2. When we say X and Y are *I.I.D.*, we mean they are independent and identically distributed.

Theorem 2.4.1 (*Linear Combinations of RVs*). *Let $\{X_i\}_{i=1}^n$ be independent random variables. If $Y = a_1X_1 + ... + a_nX_n + b$, then we have*

- $\mathbb{E}[Y] = a_1\mathbb{E}[X_1] + ... + a_n\mathbb{E}[X_n] + b$
- $Var[Y] = a_1^2 Var[X_1] + ... + a_n^2 Var[X_n]$

In particular, let $\{X_i\}_{i=1}^n$ be i.i.d., if $Y = X_1 + ... + X_n$, then we have

- $\mathbb{E}[Y] = n\mathbb{E}[X_i]$
- $Var[Y] = nVar[X_i]$

The statements regarding expectations hold even if they're ***not*** *independent.*

2.5 Bernoulli Random Variables

Definition 2.5.1 (*Bernoulli Trials*)**.** *A p-Bernoulli trail is an experiment that results in one of two outcomes: a success with probability p, or a failure with probability $1-p$.*

Definition 2.5.2 (*Bernoulli RVs*)**.** *Let X be the number of success on one p-Bernoulli trial, that is,*

$$X(\omega) = \begin{cases} 1, \text{ if } \omega \text{ is a success} \\ 0, \text{ if } \omega \text{ is a failure,} \end{cases}$$

then X is called a p-Bernoulli random variable, denoted by $X \sim Bernoulli(p)$.

Lemma 2.5.1 (*Pmf of Bernoulli RVs*)**.** *Let $X \sim Bernoulli(p)$, then*

$$P(X = x) = \begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$$

Lemma 2.5.2 (*Expectations & Variance of Bernoulli RVs*)**.** *Let $X \sim Bernoulli(p)$, then*

$$\mathbb{E}[X] = p, \; Var[X] = p(1-p).$$

Proof. $\mathbb{E}[X] = \sum\limits_{x=0,1} x \cdot P(X = x) = p$, and
$Var[X] = \sum\limits_{x=0,1} (x-\mu)^2 \cdot P(X = x) = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = p(1-p)$ □

Remark 2.5.1. When we say $X \sim Bernoulli(p)$, we essentially mean the probability distribution of X has such properties. So to prove $X \sim Bernoulli(p)$, we only need to show such properties hold for X.

2.6 Binomial Random Variables

Definition 2.6.1 (*Binomial RVs*). *Consider a sequence of independent p-Bernoulli trials. Let X be the number of successes on n trials, then X is called a (n,p)-Binomial random variable, denoted by $X \sim Binomial(n,p)$. Note that $Binomial(1,p) = Bernoulli(p)$.*

Lemma 2.6.1 (*Probability Distribution of Binomial RVs*). *Let $X \sim Binomial(n,p)$, then*

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},\ x \in \{0, 1, \dots n\}$$

Proof. range(X) $= \{0, 1, \dots n\}$, and for each $x \in \text{range}(X)$, we mean exactly x successes and $n-x$ failures, so

$$P(X = x) = \underbrace{\binom{n}{x}}_{\text{\# of ways to choose x successes}} \underbrace{p^x (1-p)^{n-x}}_{\text{by independence}}.$$

To verify it satisfies $P(\Omega) = 1$,

$$\sum_{x=0}^{n} P(X = x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1.$$

□

Theorem 2.6.1. *Let $\{X_i\}_{i=1}^n \sim Bernoulli(p)$ be independent RVs, then*

$$\sum_{i=1}^{n} X_i \sim Binomial(n,p),$$

and we can always a (n,p)-Binomial RV as a sum of n independent p-Bernoulli RVs.

Proof. range$(\sum\limits_{i=1}^{n} X_i) \in \{0, 1, \dots, n\}$, and then

$$P(\sum_{i=1}^{n} X_i = x) = \underbrace{\binom{n}{x}}_{\text{\# of ways to choose x RVs to be 1}} \underbrace{p^x (1-p)^{n-x}}_{\text{by independence}}.$$

□

Lemma 2.6.2 (*Expectations & Variance of Binomial RVs*). *Let $X \sim Binomial(n,p)$, then*

$$\mathbb{E}[X] = np,\ Var[X] = np(1-p).$$

Proof. Write X as a sum of independent $X_i \sim Bernoulli(p)$, i.e. $X = \sum\limits_{i=1}^{n} X_i$, then

$$\mathbb{E}[X_1 + X_2 + \dots X_n] = n\mathbb{E}[X_i] = np, \text{ and}$$
$$Var[X1 + X_2 + \dots X_n] = nVar[X_i] = np(1-p)$$

Of course, we could use the definition to prove it. □

2.7 Geometric Random Variables

Definition 2.7.1 (*Geometric RVs*)**.** *Consider a sequence of independent p-Bernoulli trials. Let X be the number of trials until the first success, then X is called a p-Geometric random variable, denoted by $X \sim$ Geometric(p).*

Lemma 2.7.1 (*Pmf of Geometric RVs*)**.** *Let $X \sim$ Geometric(p), then*

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

Proof. The first success happens on trial x, so we have the previous $x - 1$ trails are all failures, so by independence

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

To verify $P(\Omega) = 1$,

$$\sum_{x=1}^{\infty} P(X = x) = \sum_{x=1}^{\infty}(1-p)^{x-1}p = p\sum_{x=1}^{\infty}(1-p)^{x-1} = p \cdot \frac{1}{1-(1-p)} = 1.$$

□

Lemma 2.7.2 (*Expectations & Variance of Geometric RVs*)**.** *Let $X \sim$ Geometric(p), then*

$$\mathbb{E}[X] = \frac{1}{p},\ Var[X] = \frac{1-p}{p^2}.$$

Proof. Skipped for now. □

Lemma 2.7.3. *Let $X \sim$ Geometric(p), then*

$$\begin{aligned} &P(X \geq x) = (1-p)^{x-1}, \\ &P(X \leq x) = 1 - P(X \geq x+1) = 1 - (1-p)^x. \end{aligned}$$

Proof.

$$\begin{aligned} P(X \geq x) &= P(X = x) + P(X = x+1) + P(X = x+2) + \dots \\ &= (1-p)^{x-1}p + (1-p)^x p + (1-p)^{x+1}p + \dots \\ &= (1-p)^{x-1}p \cdot (1 + (1-p) + (1-p)^2 + \dots) \\ &= (1-p)^{x-1}p \cdot \frac{1}{1-(1-p)} \\ &= (1-p)^{x-1} \end{aligned}$$

□

Remark 2.7.1. $P(X \geq x)$ measures the probability that the first success doesn't occur on the first $x - 1$ trials.

Lemma 2.7.4 (*Memoryless Property*)**.** *Let $X \sim$ Geometric(p), then*

$$P(X > a + b | X > a) = P(X > b).$$

Proof.

$$\begin{aligned}
P(X > a + b|X > a) &= \frac{P(X > a + b)}{P(X > a)} \\
&= \frac{P(X \geq a + b + 1)}{P(X \geq a + 1)} \\
&= \frac{(1-p)^{a+b}}{(1-p)^{a}} \\
&= (1-p)^{b} \\
&= P(X \geq b + 1) \\
&= P(X > b)
\end{aligned}$$

□

Remark 2.7.2. $P(X > a + b|X > a)$ measures the probability that the first success doesn't occur on the first $a + b$ trails, given that it doesn't occur on the first a trails. So, we can define $Y \sim \textit{Geometric}(p)$ measuring the first success starting from the previous $a + 1^{\text{th}}$ trails. We need $P(Y > b)$.

2.8 Negative Binomial Random Variable

Definition 2.8.1 (*Negative Binomial RVs*)**.** *Consider a sequence of independent p-Bernoulli trials. Let X be the number of trials until the r^{th} success, then X is called a (r, p)-Negative Binomial random variable, denoted by $X \sim NegBin(r, p)$.*

Lemma 2.8.1 (*Pmf of Negative Binomial RVs*)**.** *Let $X \sim NegBin(r, p)$, then*

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r},\ x \in \{r, r+1, \dots\}.$$

Proof. We need to pick $x - 1$ successes out of the first $r - 1$ trials, the rest $x - r$ trails are failures, and then the x^{th} trial is a success. Then by independence,

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

To verify that $P(\Omega) = 1$, □

Theorem 2.8.1. *Let $\{X_i\}_{i=1}^{r} \sim Geometric(p)$ be independent RVs, then*

$$\sum_{i=1}^{r} X_i \sim NegBin(r, p).$$

Proof. □

Lemma 2.8.2 (*Expectations and Variance of Negative Binomial RVs*)**.** *Let $X \sim NegBin(r, p)$, then*

$$\mathbb{E}[X] = \frac{r}{p},\ Var[X] = \frac{r(1-p)}{p^2}.$$

Proof. Write X as a sum of independent $X_i \sim Geometric(p)$, i.e. $X = \sum_{i=1}^{n} X_i$, then

$$\mathbb{E}[X_1 + X_2 + \dots X_n] = n\mathbb{E}[X_i] = \frac{r}{p}, \text{ and}$$
$$Var[X1 + X_2 + \dots X_n] = nVar[X_i] = \frac{r(1-p)}{p^2}$$

□

2.9 Poisson Random Variables

Consider that we split a time interval into n parts with equal length. For each interval, either success or failure will occur with probability p and $1-p$. Let X be the number of successes with $\mu_X = \lambda$, where λ is fixed in advance. Clearly, $X \sim Binomial(n,p)$, so $\mu_X = \lambda = np$, then we have

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

If success or failure is allowed to occur continuously (at any times) within the time interval, that is, n approaches ∞, while maintaining $\mu_X = \lambda = np$, then

$$P(X=x) = \lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} = \lim_{n\to\infty} \binom{n}{x} (\frac{\lambda}{n})^x (1-\frac{\lambda}{n})^{n-x} = \frac{\lambda^x}{x!} e^{-\lambda}.$$

Then X is called a $Poisson(\lambda)$ random variable, where $\mu_X = \lambda$.

Definition 2.9.1 (*Poisson RVs*)**.** *Consider a continuous time period where successes occur uniformly during the time period, with an average of λ successes per period. Define the random variable X as the number of successes in that time period, then X is called a $Poisson(\lambda)$ random variable.*

Lemma 2.9.1 (*Pmf of Poisson RVs*)**.** *Let $X \sim Poisson(\lambda)$, then*

$$P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda},\ x \in \{0,1,2,\dots\}.$$

Lemma 2.9.2 (*Expectations and Variance of Poisson RVs*)**.** *Let $X \sim Poisson(\lambda)$, then*

$$\mathbb{E}[X] = \lambda,\ Var[X] = \lambda.$$

Proof. Skipped for now. □

Theorem 2.9.1. *Consider n independent $\{X_i \sim Poisson(\lambda_i)\}_{i=1}^n$ random variables, then*

$$\sum_{i=1}^n X_i \sim Poisson(\sum_{i=1}^n \lambda_i).$$

Proof. Skipped for now. □

Theorem 2.9.2 (*Poisson Approximation to the Binomial*)**.** *Let $X \sim Binomial(n,p)$, then for large n and small p, we have*

$$X \approx Poisson(np).$$

Proof. Skipped for now. □

Theorem 2.9.3. *Let $X \sim Poisson(\mu_X = \lambda)$, and each success is considered "special" with a probability of p. Let Y be the number of special success, then we have*

$$Y \sim Poisson(\mu_Y = p\lambda).$$

Proof. Skipped for now. □

Theorem 2.9.4 (*Poisson Counting Process*)**.** *Let $X \sim$ Poisson(λ) be the number of successes in one time unit with $\mu_X = \lambda$. Let Y_t be the number of successes in a time interval of length t, then we have*

$$Y_t \sim \textit{Poisson}(\mu_{Y_t} = t\lambda), \textit{ for all } t \geq 0,$$

assuming the successes occur uniformly over the time interval.

Proof. Skipped for now. □

2.10 Hypergeometric Random Variables

Definition 2.10.1 (*Hypergeometric Random Variables*)**.** *Consider a population with a finite number of elements N, where r of those elements are considered "successes". We then select n out of the N elements without replacement. Define X as the number of successes in the selection. X is called a Hypergeo(N, r, n) random variable.*

Lemma 2.10.1 (*Pmf of Hypergeometric RVs*)**.** *Let $X \sim$ Hypergeo(N, r, n), then*

$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

Proof. Skipped for now. □

Theorem 2.10.1 (*Expectations and Variance of Hypergeometric RVs*)**.** *Let $X \sim$ Hypergeo(N, r, n), then*

$$\mathbb{E}[X] = \frac{nr}{N}, \, Var[X] = n(\frac{r}{N})(\frac{N-r}{N})(\frac{N-n}{N-1}).$$

Proof. Skipped for now. □

3 Continuous Random Variables

3.1 Cumulative Distribution Function

Definition 3.1.1 (*Cumulative Distribution Function*)**.** *For any random variable X, the cumulative distribution function (cdf) of X, $F_X : \mathbb{R} \to [0,1]$, is defined as*

$$F_X(x) = P(X \leq x).$$

Lemma 3.1.1 (*Properties of cdf*)**.** *For any random variable X, the followings hold for its cdf F_X:*

- $\lim\limits_{x\to-\infty} F_X(x) = 0$,
- $\lim\limits_{x\to\infty} F_X(x) = 1$,
- F_X *is non-decreasing,*
- F_X *is right-continuous, i.e.,* $\forall\, a \in \mathbb{R} : \lim\limits_{x\to a^+} F_X(x) = F_X(a)$,
- $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$.

Lemma 3.1.2 (*Probability Masses Calculation*)**.** *Let X be any random variable, then its probability mass at a can be calculated using its cdf as follows:*

$$\begin{aligned} P(X = a) &= P(X \leq a) - \lim_{x\to a^-} P(X \leq x) \\ &= F_X(a) - \lim_{x\to a^-} F_X(x). \end{aligned}$$

3.2 Continuous Random Variables

Definition 3.2.1 (*Continuous Random Variables*)**.** *Let X be a random variable, then X is called continuous if*

$$F_X(x) \text{ is continuous for all } x.$$

Definition 3.2.2 (*Probability Density Function*)**.** *Let X be a continuous random variable, and F_X be its cdf, then it's probability density function (pdf) $f_X : \mathbb{R} \to \mathbb{R}$ is defined as*

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

In other words,

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

Proof. To see this, by the Fundamental theorem of calculus,

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

□

Lemma 3.2.1 (*Properties of pdf*)**.** *Let X be a continuous random variable, and f_X be its pdf, then we have*

- $\forall x \in \mathbb{R} : f_X(x) \geq 0,$
- $\int\limits_{-\infty}^{\infty} f_X(x)dx = 1,$
- $P(X = a) = 0,$
- $P(a < X \leq b) = \int\limits_{a}^{b} f_X(x)dx = P(a \leq X < b) = P(a < X < b) = P(a \leq X \leq b).$

Proof. Let's prove the above lemma piece by piece:

- Assume $f_X(x) < 0$ for some x, then $F_X(a^-) = \int\limits_{-\infty}^{a^-} f_X(x)dx > \int\limits_{-\infty}^{a} f_X(x)dx = F_X(a)$, which contradicts with F_X being non-decreasing,
- $\int\limits_{-\infty}^{\infty} f_X(x)dx = \lim\limits_{x\to\infty} F_X(x) = 1,$
- $P(X = a) = F_X(a) - \lim\limits_{x\to a^-} F_X(x) = 0$, since F_X is continuous,
- $P(a < X \leq b) = F_X(b) - F_X(a) = \int\limits_{a}^{b} f_X(x)dx$, by FTC.

□

Remark 3.2.1. The continuity of pdf doesn't necessarily follow that of cdf.

Definition 3.2.3 (*Percentiles*)**.** *Let X be a continuous random variable, then the $100p^{th}$ percentile $(0 < p < 1)$ of X is the value π_p s.t.*

$$F_X(\pi_p) = P(X \leq \pi_p) = p.$$

The median is defined as the 50^{th} percentile.

3.3 Expectations

Definition 3.3.1 (*Raw Moments*)**.** *Let X be a continuous random variable, then the k^{th} raw moment of X is defined as*

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

Definition 3.3.2 (*Expectations*)**.** *Let X be a continuous random variable, then the expected value of X is defined as its first raw moment, i.e.,*

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

Theorem 3.3.1 (*Compositions of Expectation*)**.** *Let g be a real-valued function of X, then*

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

Theorem 3.3.2 (*Properties of Expectations*)**.** *Let $\{g_i\}_{i=1}^n$ be real-valued functions of X, and $\{c_i\}_{i=1}^n$ be constants. Then, we have*

- $\mathbb{E}[c] = c,$
- $\mathbb{E}[cg(X)] = cE[g(X)],$
- $\mathbb{E}[g_1(X) + ... + g_n(X)] = \mathbb{E}[g_1(X)] + ... + \mathbb{E}[g_n(X)].$

Theorem 3.3.3 (*Scaling & Shifting on Expectations*)**.** *Let a, b be constants, we have the following:*

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Theorem 3.3.4 (*Linear Combinations of RVs*)**.** *Let $\{X_i\}_{i=1}^n$ be independent random variables. If $Y = a_1X_1 + ... + a_nX_n + b$, then we have*

$$\mathbb{E}[Y] = a_1\mathbb{E}[X_1] + ... + a_n\mathbb{E}[X_n] + b.$$

In particular, let $\{X_i\}_{i=1}^n$ be i.i.d., if $Y = X_1 + ... + X_n$, then we have

$$\mathbb{E}[Y] = n\mathbb{E}[X_i].$$

*The statements involving expectations hold even if they're **not** independent.*

3.4 Variance

Definition 3.4.1 (*Central Moments*)**.** *Let X be a continuous random variable, then the k^{th} central moment of X is defined as*

$$\mathbb{E}[(X-\mu_X)^k] = \int\limits_{-\infty}^{\infty} (x-\mu_X)^k f_X(x)dx.$$

Definition 3.4.2 (*Variance*)**.** *Let X be a continuous random variable, then the variance of X is defined as its second central moment, i.e.,*

$$\sigma_X^2 = Var[X] = \int\limits_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx.$$

Theorem 3.4.1 (*Compositions of Variance*)**.** *Let g be a real-valued function of X, then*

$$Var[g(X)] = \int\limits_{-\infty}^{\infty} (g(x)-\mathbb{E}[g(X)]) f_X(x)dx.$$

Theorem 3.4.2 (*Variance Using Expectations*)**.** *We can calculate the variance of X by*

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

More generally, $Var[g(X)] = \mathbb{E}[g(X)^2] - \mathbb{E}[g(X)]^2$.

Theorem 3.4.3 (*Shifting & Scaling on Variance*)**.** *Let a, b be constants, we have the following:*

$$Var[aX+b] = a^2 Var[X].$$

In particular, $Var[aX] = a^2 Var[X]$, and $Var[b] = 0$.

Theorem 3.4.4 (*Linear Combinations of RVs*)**.** *Let $\{X_i\}_{i=1}^n$ be independent random variables. If $Y = a_1X_1 + ... + a_nX_n + b$, then we have*

$$Var[Y] = a_1^2 Var[X_1] + ... + a_n^2 Var[X_n]$$

In particular, let $\{X_i\}_{i=1}^n$ be i.i.d., if $Y = X_1 + ... + X_n$, then we have

$$Var[Y] = nVar[X_i].$$

3.5 Moment Generating Functions

Definition 3.5.1 (*Moment Generating Functions*)**.** *Let X be a random variable, then the moment generating function (mgf) is defined as*

$$m_X(t) = \mathbb{E}[e^{tX}].$$

Theorem 3.5.1. *Let X be a random variable. If $m_X(t)$ exists, then it is unique, that is,*

$$m_X(t) = m_Y(t) \Leftrightarrow X = Y,$$

which means they share the same distribution.

Theorem 3.5.2. *Let X be a random variable. If $m_X(t)$ exists, then for any $k \in \mathbb{N}_{>0}$, then*

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

Theorem 3.5.3. *Let $\{X_i\}_{i=1}^n$ be independent random variables with mgf's $\{m_{X_i}(t)\}_{i=1}^n$. Let $Y = \sum\limits_{i=1}^n X_i$, then we have*

$$m_Y(t) = \prod_{i=1}^n m_{X_i}(t).$$

In particular, if they're i.i.d., $m_Y(t) = [m_{X_i}(t)]^n$.

Lemma 3.5.1. *Let $X \sim$ Binomial(n, p), then we have*

$$m_X(t) = (pe^t + 1 - p)^n.$$

3.6 Uniform Distribution

Definition 3.6.1 (*Uniform Distribution*)**.** *Let X be a random variable, then X has a uniform distribution with parameters $\theta_1 < \theta_2$, if its pdf has the form:*

$$f_X(x) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \textit{if } \theta_1 \le x \le \theta_2 \\ 0, & \textit{otherwise} \end{cases}$$

then we write $X \sim \textit{Uniform}(\theta_1, \theta_2)$.

Remark 3.6.1. Uniform distribution is similar to "equally likely" in some sense - as long as the intervals are of equal length, then the probability density will be the same, that is,

$$|b - a| = |d - c| \Rightarrow P(a \le X \le b) = P(c \le X \le d).$$

Informally, we could consider each continuous point has the same probability. If we say for example "a point is chosen **at random** on the interval $[0, 2]$", we mean it follows $\textit{Uniform}(0, 2)$.

Lemma 3.6.1 (*cdf of Uniform Distribution*)**.** *Let $X \sim \textit{Uniform}(\theta_1, \theta_2)$, then*

$$F_X(x) = \begin{cases} 0, & \textit{if } x < \theta_1 \\ \frac{x - \theta_1}{\theta_2 - \theta_1}, & \textit{if } \theta_1 \le x \le \theta_2 \\ 1, & \textit{if } x > \theta_2 \end{cases}$$

Lemma 3.6.2. *Let $X \sim \textit{Uniform}(\theta_1, \theta_2)$, then*

$$P(a \le X \le b) = \frac{b - a}{\theta_2 - \theta_1}.$$

Theorem 3.6.1 (*Expectations & Variance of Uniform Distribution*)**.** *Let $X \sim \textit{Uniform}(\theta_1, \theta_2)$, then*

$$\mathbb{E}[X] = \frac{\theta_1 + \theta_2}{2}, \textit{ and } Var[X] = \frac{(\theta_2 - \theta_1)^2}{12}.$$

Proof. Skipped for now. □

Lemma 3.6.3 (*mgf of Uniform Distribution*)**.** *Let $X \sim \textit{Uniform}(\theta_1, \theta_2)$, then*

$$m_X(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}.$$

Proof.

$$m_X(t) = \mathbb{E}[e^{tX}] = \int\limits_{\theta_1}^{\theta_2} e^{tx} \frac{1}{\theta_2 - \theta_1} dx = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}.$$

□

Theorem 3.6.2 (*Linear Function of Uniform Distribution*)**.** *Let $X \sim \textit{Uniform}(\theta_1, \theta_2)$, $a > 0$, and b be constants, then*

$$Y = aX + b \sim \textit{Uniform}(a\theta_1 + b, a\theta_2 + b).$$

Proof. We will prove m_Y is the mgf of a uniform RV w.r.t. some parameters, and then by the uniqueness of mgf, $Y \sim$ *Uniform* w.r.t. those parameters.

$$m_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}[e^{taY+tb}] = e^{tb}\mathbb{E}[e^{taY}] = e^{tb}m_Y(ta),$$

which holds for any Y. Then using the mgf of uniform RV, we have

$$m_Y(t) = etb\frac{e^{ta\theta_2} - e^{ta\theta_1}}{ta(\theta_2 - \theta_1)} = \frac{e^{t(a\theta_2+b)} - e^{t(a\theta_1+b)}}{t(a\theta_2 - a\theta_1)} = \frac{e^{t(a\theta_2+b)} - e^{t(a\theta_1+b)}}{t((a\theta_2 + b) - (a\theta_1 + b))},$$

We then get to the mgf of *Uniform*$(a\theta_1 + b, a\theta_2 + b)$ if $a > 0$, i.e., the first parameter is smaller than the second one. If $a < 0$, the statement holds by swapping the parameters. □

Remark 3.6.2. We could prove by showing that their cdf or pdf are equal.

3.7 Normal Distribution

Definition 3.7.1 (*Normal Distribution*)**.** *A random variable X is said to have a normal probabilitiy distribution with $\sigma^2 > 0$ and $\mu \in \mathbb{R}$, denoted by $X \sim Normal(\mu, \sigma^2)$ if its pdf for all $x \in \mathbb{R}$ has the form:*

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

Definition 3.7.2 (*Standard Normal Distribution*)**.** *A random variable Z is said to have a standard normal distribution if*

$$Z \sim Normal(0, 1).$$

Remark 3.7.1. There's no closed form formula for $P(a \leq X \leq b)$, we'll see how to handle this shortly.

Theorem 3.7.1 (*Expectations & Variance of Normal Distribution*)**.** *Let $X \sim Normal(\mu, \sigma^2)$, then*

$$\mathbb{E}[X] = \mu, \text{ and } Var[X] = \sigma^2.$$

Proof. Skipped for now. □

Lemma 3.7.1. *Let $X \sim Normal(\mu, \sigma^2)$, and $Y = aX + b$, then*

$$Y \sim Normal(a\mu + b, a^2\sigma^2).$$

Proof. Skipped for now. □

Lemma 3.7.2. *Let $X \sim Normal(\mu, \sigma^2)$, then*

$$Z = \frac{X - \mu}{\sigma} \sim Normal(0, 1),$$

and therefore,

$$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}).$$

Remark 3.7.2. We provide a z-table for standard normal distribution, and we can compute any normal distribution based on this by standardizing first.

Theorem 3.7.2. *Let $\{X_i\}_{i=1}^n \sim \mathcal{N}(\mu_i, \sigma_i^2)$ be independent random variables, then*

$$\sum_{i=1}^n aX_i \sim \mathcal{N}(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2).$$

In particular, if $\{X_i\}_{i=1}^n \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2).$$

3.8 Gamma Distribution

Definition 3.8.1. *A random variable X is said to have a Gamma distribution with $\alpha > 0$ and $\beta > 0$, denoted by $X \sim Gamma(\alpha, \beta)$ if its pdf has the form:*

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, & \textit{if } x > 0, \\ 0, & \textit{otherwise,} \end{cases}$$

where $\Gamma(\alpha) = \int\limits_0^\infty x^{\alpha-1} e^{-x} dy$.

Lemma 3.8.1 (*Properties of* Γ)**.** *The followings holds for $\Gamma(\alpha)$:*

- $\Gamma(\alpha) > 0$ *for all* $\alpha > 0$,
- $\Gamma(1) = 1$,
- $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ *for all* $\alpha > 1$,
- $\Gamma(n) = (n - 1)!$ *for all* $n \in \mathbb{N}_+$,
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$,
- $\int\limits_0^\infty x^{\alpha-1} e^{\frac{x}{\beta}} dx = \beta^\alpha \Gamma(\alpha)$.

Lemma 3.8.2. *Let $X \sim Gamma(\alpha, \beta)$, then its mgf has the form:*

$$m_X(t) = (1 - \beta t)^{-\alpha}.$$

Theorem 3.8.1. *Let $X \sim Gamma(\alpha, \beta)$, then*

$$\mathbb{E}[X] = \alpha\beta, \ \textit{and } Var[X] = \alpha\beta^2.$$

3.9 Chi-Square Distribution

Definition 3.9.1 (*Chi-Square Distribution*)**.** *A random variable X is said to have a Chi-Square distribution with $v > 0$ denoted by $X \sim \chi_v^2$ if*

$$X \sim Gamma(\alpha = \frac{v}{2}, \beta = 2),$$

where v is called the degree of freedom.

Theorem 3.9.1 (*Expectation & Variance of Chi-Square Distribution*)**.** *Let $X \sim \chi_v^2$, then*

$$\mathbb{E}[X] = v, \ \textit{and} \ Var[X] = 2v.$$

Lemma 3.9.1. *Let $X \sim \mathcal{N}(0, 1)$, then*

$$X^2 \sim \chi_{v=1}^2.$$

3.10 Exponential Distribution

Definition 3.10.1. *A random variable X is said to have a Exponential distribution with $\beta > 0$, denoted by $X \sim Exp(\beta)$ if*

$$X \sim Gamma(\alpha = 1, \beta),$$

and hence its pdf has the form:

$$f_X(x) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 3.10.1 (*cdf of Exponential Distribution*)**.** *Let $X \sim Exp(\beta)$, then it pdf has the form:*

$$F_X(x) = \begin{cases} 1 - e^{-\frac{x}{\beta}}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

and therefore, $P(a \leq X \leq b) = e^{-\frac{a}{\beta}} - e^{-\frac{b}{\beta}}$.

Lemma 3.10.2 (*mgf of Exponential Distribution*)**.** *Let $X \sim Exp(\beta)$, then its mgf has the form:*

$$m_X(t) = (1 - \beta t)^{-1}.$$

Theorem 3.10.1 (*Expectations & Variance of Exponential Distribution*)**.** *Let $X \sim Exp(\beta)$, then*

$$\mathbb{E}[X] = \beta, \text{ and } Var[X] = \beta^2.$$

Lemma 3.10.3 (*Memoryless Property*)**.** *Let $X \sim Exp(\beta)$. Then for any $a, b > 0$, we have*

$$P(X > a + b | X > a) = P(X > b).$$

Theorem 3.10.2. *Let $\{X_i\}_{i=1}^n \sim Exp(\beta)$ be independent. Then*

$$\sum_{i=1}^n X_i \sim Gamma(\alpha = n, \beta).$$

Theorem 3.10.3. *Let $X \sim Exp(\beta)$ and $Y = aX$ where $a > 0$, then*

$$Y \sim Exp(a\beta).$$

Note that it doesn't hold if $Y = aX + b$.

3.11 Beta Distribution

Definition 3.11.1. *A random variable X is said to have a Beta distribution with $\alpha, \beta > 0$, denoted by $X \sim Beta(\alpha, \beta)$ if its pdf has the form:*

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & \textit{if } 0 < x < 1, \\ 0, & \textit{otherwise} \end{cases}$$

where $\Gamma(\alpha) = \int\limits_0^\infty x^{\alpha-1} e^{-x} dy$.

Theorem 3.11.1 (*Expectation & Variance of Beat Distribution*)**.** *Let $X \sim Beta(\alpha, \beta)$, then*

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \textit{ and } Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

4 Multivariate Distribution

4.1 Discrete Case

Definition 4.1.1 (*Joint pmf*)**.** *Let X, Y be discrete random variables. Then the joint probability distribution of X and Y is the set of probabilities $\{P(X = x, Y = y)\}_{all\ x,y}$, and $P_{X,Y}$ is called the joint pmf.*

Lemma 4.1.1. *The above joint pmf $P_{X,Y}$ satisfies the followings:*

- $\forall x, y : P(x, y) \in [0, 1]$,
- $\sum\limits_{all\ x} \sum\limits_{all\ y} P(x, y) = 1.$

Definition 4.1.2 (*Marginal pmf*)**.** *Let X, Y be discrete random variables with joint pmf $P_{X,Y}(x, y)$, then the marginal pmf of X is computed as*

$$P_X(x) = \sum_{all\ y} P_{X,Y}(x, y).$$

Definition 4.1.3 (*Conditional pmf*)**.** *Let X, Y be discrete random variables with joint pmf $P_{X,Y}(x, y)$, and marginal pmf f_X, f_Y. Then the conditional pmf of X given $Y = y$ is*

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x, y)}{P_Y(y)},$$

provided that $P_Y(y) > 0$.

4.2 Continuous Case

Definition 4.2.1 (*Joint pdf*)**.** *Let X, Y be continuous random variables with joint cdf $F(x,y) := P(X \leq x, Y \leq y)$. If there exists a non-negative function $f(x,y)$ s.t.*

$$F(x,y) = \int\limits_{-\infty}^{x} \int\limits_{\infty}^{y} f(s,t)dsdt$$

for all $x, y \in \mathbb{R}$, then X, Y are called jointly continuous random variables, and $f_{X,Y}(x,y)$ is called the joint pdf.

Lemma 4.2.1. *Let X, Y be continuous random variables with $F(x,y)$. By definition, we have*

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y).$$

Lemma 4.2.2. *The above joint pdf $f(x,y)$ safisfies the followings:*

- $\forall x, y \in \mathbb{R} : f(x,y) \geq 0$,
- $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x,y)dA = 1$,
- $P((X,Y) \in \mathcal{D}) = \iint\limits_{\mathcal{D}} f(x,y)dA.$

Remark 4.2.1. The probability previously is the area under the curve, while now is the volume.

Definition 4.2.2 (*Marginal pdf*)**.** *Let X, Y be continuous random variables with joint pdf $f(x,y)$, then the marginal pdf of X is computed as*

$$f_X(x) = \int\limits_{-\infty}^{\infty} f(x,y)dy.$$

Definition 4.2.3 (*Conditional cdf*)**.** *Let X, Y be continuous random variables with joint pdf $f(x,y)$ and marginal pdf f_X, f_Y, then the conditional pdf of X given $Y = y$ is computed as*

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

provided that $f_Y(y) > 0$.

Remark 4.2.2. Some calculations (continuous random variables):

- $P(X \leq a | Y = b) = \int_{-\infty}^{a} f_{X|Y=b}(x)dx$,
- $P(X \leq a | Y \leq b) = \frac{P(X \leq a, Y \leq b)}{P(Y \leq b)} = \frac{\int_{-\infty}^{a} \int_{-\infty}^{b} f(x,y)dA}{\int_{-\infty}^{b} f_Y(y)dy}$.

4.3 Independence

Definition 4.3.1 (*Independence of RVs*)**.** *Let X, Y be random variables with cdfs F_X, F_Y and their joint cdf $F_{X,Y}$. Then X and Y are independent if for all x, y,*

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

Lemma 4.3.1 (*Independence of Discrete RVs*)**.** *Let X, Y be discrete random variables, then X and Y are independent if and only if any of the followings hold:*

- $P(X = x, Y = y) = P(X = x)P(Y = y)$,
- $P(X = x|Y = y) = P(X = x)$,
- $P(Y = y|X = x) = P(Y = y)$.

Lemma 4.3.2 (*Independence of Continuous RVs*)**.** *Let X, Y be continuous random variables, then X and Y are independent if and only if any of the followings hold:*

- $f(x, y) = f_X(x)f_Y(y)$,
- $f(x|y) = f_X(x)$,
- $f(y|x) = f_Y(y)$.

Theorem 4.3.1. *Let X, Y be continuous random variables with joint pdf $f(x, y)$ where*

- $\forall a \leq x \leq b, \forall c \leq y \leq d, f(x, y) > 0$, *for some* $a, b, c, d \in \overline{\mathbb{R}}$,
- $f(x, y) = 0$ *elsewhere,*

then X and Y are independent iff

$$f(x, y) = g(x)h(y),$$

for some function g, h.

Remark 4.3.1. Note that when $f(x, y) > 0$ and X depends on Y, then X and Y are dependent.

4.4 Expectations & Variance

Definition 4.4.1 (*Expectations of Discrete RVs*)**.** *Let X, Y be discrete random variables with the joint pmf $P(x,y)$, then the expectation of $g(X,Y)$ is*

$$\mathbb{E}[g(X,Y)] = \sum_{all\ x} \sum_{all\ y} g(x,y)P(x,y).$$

Definition 4.4.2 (*Expectations of Continuous RVs*)**.** *Let X, Y be continuous random variables with the joint pdf $f(x,y)$, then the expectation of $g(X,Y)$ is*

$$\mathbb{E}[g(X,Y)] = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} g(x,y)f(x,y)dA.$$

Theorem 4.4.1 (*Variance*)**.** *Let X, Y be continuous random variables, then the variance of $g(X,Y)$ is*

$$Var[g(X,Y)] = \mathbb{E}[g(X,Y)^2] - \mathbb{E}[g(X,Y)]^2.$$

Theorem 4.4.2. *Let $\{X_i\}_{i=1}^n$ be random variables, then*

- $\mathbb{E}[c] = c$
- $\mathbb{E}[cg(X_1, ..., X_n)] = c\mathbb{E}[g(X_1, ..., X_n)]$
- $\mathbb{E}[\sum_{i=1}^m g_i(X_1, ..., X_n)] = \sum_{i=1}^m \mathbb{E}[g_i(X_1, ..., X_n)]$

Theorem 4.4.3. *Let X, Y be independent random variables, then*

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)],$$

provided that the expectations exist. This generally doesn't hold for variance.

Theorem 4.4.4. *Let X, Y be independent random variables, then*

$$Var[X - Y] = Var[X] + Var[-Y] = Var[X] + Var[Y].$$

4.5 Conditional Expectations & Variance

Definition 4.5.1 (*Conditional Expectations of Discrete RVs*)**.** *Let X, Y be discrete random variables with the joint pmf $P(x, y)$, then the conditional expectations of $g(X)$ on $Y = y$ is defined as*

$$\mathbb{E}[g(X)|Y = y] = \sum_{all\ x} g(x)P_{X|Y=y}(x).$$

Definition 4.5.2 (*Conditional Expectations of Continuous RVs*)**.** *Let X, Y be continuous random variables with the joint pdf $f(x, y)$, then the conditional expectations of $g(X)$ on $Y = y$ is defined as*

$$\mathbb{E}[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x)f_{X|Y=y}(x)dx.$$

Lemma 4.5.1 (*Condition Variance*)**.** *Let X, Y be random variables, then the conditional expectations of $g(X)$ on $Y = y$ is defined as*

$$Var[g(X)|Y = y] = \mathbb{E}[g(X)^2|Y = y] - \mathbb{E}[g(X)|Y = y]^2.$$

Lemma 4.5.2 (*Independence*)**.** *Let X, Y be independent random variables, then we have*

$$\mathbb{E}[g(X)|Y = y] = \mathbb{E}[g(X)], \text{ and } Var[g(X)|Y = y] = Var[g(X)].$$

Lemma 4.5.3. *Let X, Y be random variables, then*

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]],$$

and regarding the variance,

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

where $Var[X|Y] = \mathbb{E}[X^2|Y] - \mathbb{E}[X|Y]^2$, and $Var[\mathbb{E}[X|Y]] = \mathbb{E}[\mathbb{E}[X|Y]^2] - \mathbb{E}[\mathbb{E}[X|Y]]^2$.

4.6 Covariance

Definition 4.6.1 (*Covariance*)**.** *Let X, Y be random variables, then the covariance of X, Y is defined as*

$$\begin{aligned} Cov[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \end{aligned}$$

Lemma 4.6.1. *Let X be a random variable, then*

- *$Cov[X, X] = Var[X]$,*
- *$Cov[c, X] = 0$, for some constant c,*
- *If X, Y are independent, then $Cov[X, Y] = 0$, while the inverse isn't always true.*

Lemma 4.6.2. *Let X, Y be random variables, then*

$$Cov[aX + b, cY + d] = acCov[X, Y].$$

Lemma 4.6.3. *Let X, Y be random variables, then*

$$Var[aX + bY + c] = a^2Var[X] + b^2Var[Y] + 2abCov[X, Y],$$

in particular, if X, Y are independent, then

$$Var[aX + bY + c] = a^2Var[X] + b^2Var[Y].$$