Probability Theory

Yanze Song

October 24, 2023

1 What are Probability & Statistics?

Statistics can be described as the science of interpreting and drawing conclusions from data. Populations are generally too large, and therefore it will be unrealistic to calculate the parameters of the populations directly. However, we can somehow sample some data and calculate the statistics of it. The process of drawing conclusions about populations (parameters) based on the data collected in samples (statistics) is called *Statistical Inference.*

Naturally, one may ask what the statistics imply the parameters, for example, how close they are. *Probability theory* will answer this kind of questions: it will help us measure the error in our estimation and how confident we can be in our inferences.

2 Probability Basics

2.1 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.

2.2 Set Theory

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

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

2.3 Probability Axioms

Probability measures the likelihood of an event. In everyday life, when saying *I think*, *Probably*, *Most likely*, etc., we are using probability, and these are called *Subjective Probability.* In the mathematical world, we need to define probability to quantify the likelihood of an event (based on some assumptions).

Definition 2.3.1 (*Probability Axioms*)**.** With respect to 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 2.3.1 (*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, $P(\bigcup\limits_{i=1}^{n} A_i) = \sum\limits_{i=1}^{n} P(A_i)$.

2.4 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 (easily proved). 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.

Example 2.4.1. Consider a random experiment as rolling a 6-sided die, and sample points are defined as the face showing up when the die stops, and hence, $S = \{1, 2, ..., 6\}$. The probability model should satisfy

$$P(\text{the roll is } i) = P(i) = \begin{cases} p_i, & \text{if } i \in \{1, 2, ..., 6\} \\ 0, & \text{otherwise,} \end{cases}$$

so that $\forall i \in \{1, 2, ..., 6\} : p_i \in [0, 1]$, and $\sum\limits_{i=1}^{6} p_i = 6$. Note that the above probability model gives no information of the exact probability of each sample points, there are infinite many of them. Now consider an event A = roll an even number, $P(A) = P(\{2, 4, 6\}) = p_2 + p_4 + p_6$.

Theorem 2.4.1 (*Equiprobable Sample Points*)**.** Given Ω with $|\Omega| = N$, assume all sample points are equiprobable, that is, 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.

2.5 Counting Methods

Theorem 2.5.1 (*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 2.5.2 (*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 2.5.3 (*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 2.5.4 (*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

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

Example 2.5.1. Consider a game that consists dealing out four hands of three cards from a deck of twelve cards. The deck contains the four Aces, the four Kings, and the 8_S, 9_H, 8_C, and 9_D. Then answer the following questions:

(1) What's the probability that every player has an ace?

4 Aces, 4 players, so $\binom{4}{1\ 1\ 1\ 1}$ ways of assigning Aces. Then 8 cards left, each player needs 2 cards, no constraints, so $\binom{8}{2\ 2\ 2\ 2}$. In total,

$$\binom{4}{1\ 1\ 1\ 1} \cdot \binom{8}{2\ 2\ 2\ 2}.$$

(2) What's the probability that exactly 3 players have an ace?

The only possible case is $2, 1, 1, 0$ Aces. $\binom{4}{1}$ to choose which player gets 2 Aces, and $\binom{3}{2}$ to choose which two players get 1 Ace. After fixing the players, $\binom{4}{2\ 1\ 1\ 0}$ ways to assign Aces, and $\binom{8}{1\ 2\ 2\ 3}$ ways to assign the rest. In total,

$$\binom{4}{1} \cdot \binom{3}{2} \cdot \binom{4}{2\ 1\ 1\ 0} \cdot \binom{8}{1\ 2\ 2\ 3}.$$

(3) What's the probability that exactly 2 players have an ace?

Two possible cases: $3, 1, 0, 0$ or $2, 2, 0, 0$ Aces. Same as above:

$$\binom{4}{1} \cdot \binom{3}{1} \cdot \binom{4}{3\ 1\ 0\ 0} \cdot \binom{8}{0\ 2\ 3\ 3} + \binom{4}{2} \cdot \binom{4}{2\ 2\ 0\ 0} \cdot \binom{8}{1\ 1\ 3\ 3}.$$

Why we do $\binom{4}{2}$ instead of $\binom{4}{1} \cdot \binom{3}{1}$? Each of them has 2 Aces, if we consider order there, then it will over-count because $\binom{4}{2\ 2\ 0\ 0}$ will consider the order as well. Or equivalently, we can divide the term by 2!.

Example 2.5.2. A quick example to compare with the above: 6 players, 3 Forwards, 2 Defensemen, 1 Goalie, $\binom{6}{3\ 2\ 1}$ ways. No need to *fix the players*.

2.6 Conditional Probability

Definition 2.6.1 (*Conditional Probability*)**.** The *conditional probability* of event A, given that event B has occurred, is defined as

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

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 2.6.2 (*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 2.6.1 (*Total Probability*)**.** Let $\{E_i\}_{i=1}^n$ be a partition of the sample space S, then for any event A, we have

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

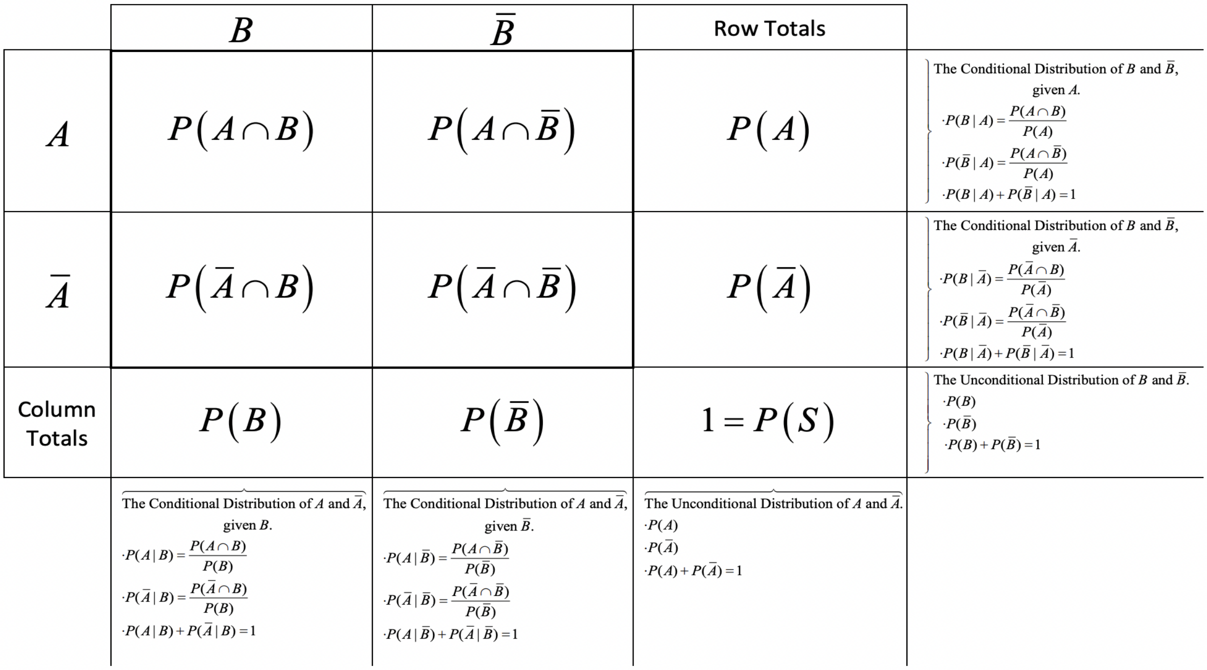
More specifically, $P(A) = P(A \cap B) + P(A \cap \overline{B})$.

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



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

2.7 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 2.7.1 (*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 2.7.1. *A and B are independent if and only if*

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

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 2.7.1 (*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 2.7.2 (*Total Probability: Revisit*)**.** 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 2.7.1. 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 2.7.3 (*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

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

Example 2.7.1. A room of n students all throw their OneCards in a hat, and then randomly draw a OneCard from the hat. If a student selects their own OneCard, then there's a match. What's the probability at least one student has a match?

Let $E_i = i^{th}$ student has a match, then

$$\begin{aligned}
P(\bigcup_{i=1}^{n} E_i) &= \sum_{i=1}^{n} P(E_i) - \sum_{i\neq j} P(E_i \cap E_j) + \sum_{i\neq j\neq k} P(E_i \cap E_j \cap E_k) - ... \\
&= n \cdot \frac{1}{n} - \binom{n}{2} P(E_i|E_j) \cdot P(E_j) + \binom{n}{3} P(E_i|E_j \cap E_k) \cdot P(E_j|E_k) \cdot P(E_k) - ... \\
&= \binom{n}{1}\frac{1}{n} - \binom{n}{2}\frac{1}{n-1} \cdot \frac{1}{n} + \binom{n}{3}\frac{1}{n-2} \cdot \frac{1}{n-1} \cdot \frac{1}{n} - ... \\
&= \binom{n}{1}\frac{1}{P_1^n} - \binom{n}{2}\frac{1}{P_2^n} + \binom{n}{3}\frac{1}{P_3^n} - ... \\
&= \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} \frac{1}{P_i^n} \\
&= \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{(n-i)! \cdot i!} \cdot \frac{(n-i)!}{n!} \\
&= \sum_{i=1}^{n} (-1)^{i+1} \frac{1}{i!}.
\end{aligned}$$

3 Discrete Random Variables

3.1 Discrete Random Variables Basics

Definition 3.1.1 (*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, $X(\omega) \in \mathbb{R}, \forall \omega \in \Omega$. X is *discrete* if range(X) is countable. Here's a common notation:

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

Remark 3.1.1. With respect to 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 are both representatives of a set of sample points in essence.

Definition 3.1.2 (*Probability Distribution / Probability Mass Function*)**.** Let X be a discrete random variable. The *probability distribution* of X is the set of probabilities $\{P(X = x)\}_{\text{all } x}$. $P(X = \cdot)$ is called the *probability mass function* (*pmf*) of X.

Remark 3.1.2. A few notations:

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

Lemma 3.1.1 (*Immediate Results*)**.** For any pmf $P_X(\cdot)$, the followings must be true:

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

Example 3.1.1. Consider flipping a coin twice, and define the random variable X to be the number of heads. Find the probability distribution of X.

range$(X) = \{0, 1, 2\}$, and hence

$$P(X = 0) = \frac{1}{4}, P(X = 1) = \frac{1}{2}, P(X = 2) = \frac{1}{4}.$$

Or equivalently,

$$P(X = x) = \underbrace{\binom{2}{x}}_{\text{\# of ways of having } x \text{ heads}} \cdot \underbrace{(\frac{1}{2})^x}_{\text{prob of having } x \text{ heads}} \cdot \underbrace{(1 - \frac{1}{2})^{3-x}}_{\text{prob of having } 3-x \text{ tails}},$$

assuming the coin is fair and all flips are independent. Later on, we'll see that X is called the *binomial random variable.*

Example 3.1.2. Consider flipping a coin until you observe your first head, and define Y to be the number of flips. Find the probability distribution of Y.

$\Omega = \{H, TH, TTH, TTTH, ...\}$, range$(Y) = \{1, 2, 3, ...\}$, and hence

$$P(Y = y) = \underbrace{(\frac{1}{2})^{y-1}}_{\text{prob of having } y-1 \text{ tails}} \cdot \underbrace{\frac{1}{2}}_{\text{prob of having 1 head}},$$

assuming the coin is fair and all flips are independent. Later on, we'll see that Y is called the *geometric random variable.*

Remark 3.1.3. Probability distribution of random variables can always be found **pointwisely** given enough information. For some specific kinds of random variables (e.g. the above two), a closed form formula exists.

Example 3.1.3. A group of four objects contains two defectives. The tester stops testing once she tests the second defective. Let X be the number of tests on which the second defective is tested. Find the probability distribution of X, assuming equally likely.

$\Omega = \{DDGG, DGDG, GDDG, DGGD, GDGD, GGDD\}$, and range$(X) = \{2, 3, 4\}$, and hence

$$P(X = 2) = P(\{DDGG\}) = \frac{1}{6},$$
$$P(X = 3) = P(\{DGDG, GDDG\}) = \frac{1}{3},$$
$$P(X = 4) = P(\{DGGD, GDGD, GGDD\}) = \frac{1}{2}.$$

Suppose testing an object costs \$2, and repairing an object costs \$4. Let Y be the total cost. Find the probability distribution of Y.

$Y = 2X + 8$, so range$(Y) = \{12, 14, 16\}$. Then

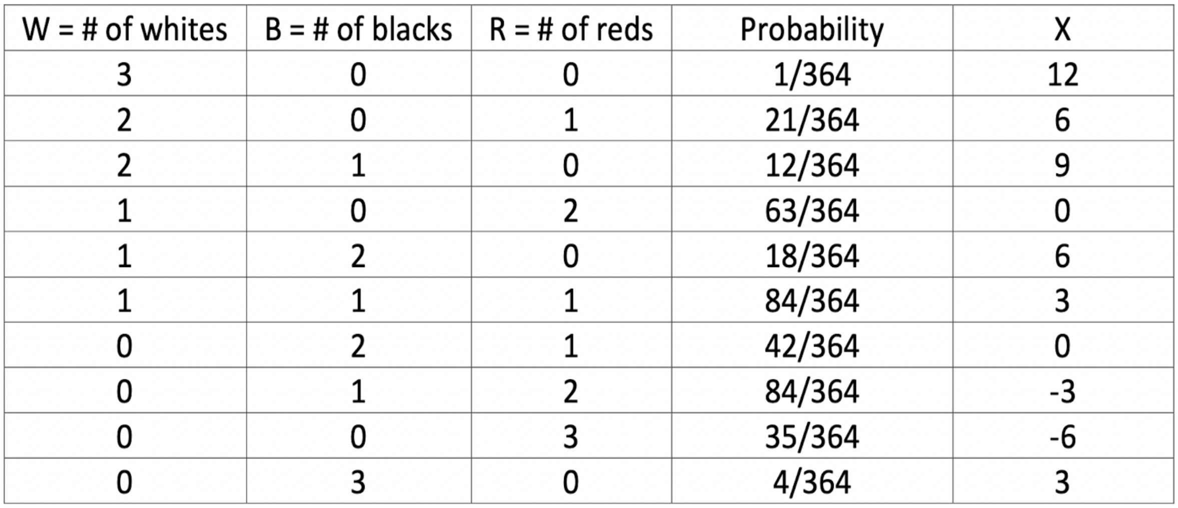
$$\begin{aligned}
P(Y = 12) &= P(2X + 8 = 12) = P(X = 2) = \frac{1}{6}, \\
P(Y = 14) &= P(X = 3) = \frac{1}{3}, \\
P(Y = 16) &= P(X = 4) = \frac{1}{2}.
\end{aligned}$$

Example 3.1.4. Randomly select 3 balls out of 3 white balls, 4 black balls, and 7 red balls. Win \$4 for each red, \$1 for each black, and lose \$2 for each red. Let X be the total winning of the game. Find the probability distribution of X.

It's not easy to find range(X) directly, so we need some intermediate RVs. Let W, B, and R be the number of each color chosen, then we have $X = 4W + B - 2R$. Then

$$P(W = x, B = b, R = r) = \frac{\binom{3}{w}\binom{4}{b}\binom{7}{r}}{\binom{14}{3}}, \forall x + b + r = 3.$$

We can find the probability distribution of (W, B, R) as below:



W = # of whites	B = # of blacks	R = # of reds	Probability	X
3	0	0	1/364	12
2	0	1	21/364	6
2	1	0	12/364	9
1	0	2	63/364	0
1	2	0	18/364	6
1	1	1	84/364	3
0	2	1	42/364	0
0	1	2	84/364	-3
0	0	3	35/364	-6
0	3	0	4/364	3

Hence, we can find that range$(X) = \{-3, 0, 3, 6, 9, 12\}$. We won't list out the entire distribution but only when $X = 3$:

$$\begin{aligned}
P(X = 3) &= P(W = 1, B = 1, R = 1 \cup W = 0, B = 3, R = 0) \\
&= P(W = 1, B = 1, R = 1) + P(W = 0, B = 3, R = 0) \\
&= \frac{88}{364},
\end{aligned}$$

where the first equality holds since they both represent the same set of sample points essentially.

Remark 3.1.4. Recall that $RV = \#$ is nothing but a set of sample points, so no matter it is the intermediate RV, e.g. $(W, B, R) = (1, 1, 1)$, or the resulting RV, e.g. $X = 3$, they all represent sets of sample points out of the same sample space. Regarding the previous example, what's special there is that each (W, B, R) represents exactly 1 sample point.

3.2 Expectations

Definition 3.2.1 (*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

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

Definition 3.2.2 (*Expectations*)**.** The *expected value* or *mean* of X is defined as the first raw moment of X:

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

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

Theorem 3.2.1 (*Composition on Expectations*)**.** Let g be a real-valued function of X, then

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

Theorem 3.2.2 (*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 3.2.3 (*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:

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

□

3.3 Variance

Definition 3.3.1 (*Central Moments*)**.** The k-th *central moment*, or the *moment about the mean* of X is defined as

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

Remark 3.3.1 (*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 3.3.2 (*Variance*)**.** The *variance* of X is defined as the second central moment of X:

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

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

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

Theorem 3.3.2 (*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 3.3.3 (*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 3.3.2. It makes sense that the shifting (b) doesn't affect the variability, because the whole distribution shifts together and the shape is maintained.

Remark 3.3.3. If $\mathbb{E}[X]$ and $Var[X]$ are known, and Y is linear on X, i.e. $Y = aX + b$, then $\mathbb{E}[Y]$ and $Var[Y]$ can be calculated without the distribution of Y.

Definition 3.3.3 (*Standard Deviation*)**.** The *standard deviation* of X is defined as

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

3.4 Expectations & Variance Practice

Example 3.4.1. Let X be a random variable with the following distribution:

$$P(X=x)=\begin{cases}\frac{1}{6}, \text{ if } x=2\\ \frac{1}{3}, \text{ if } x=3\\ \frac{1}{2}, \text{ if } x=4\end{cases}$$

Define $Y=2X+8$. Find the expected value and the variance of X and Y.

$$\mu_X=\mathbb{E}[X]=\sum_{x=2}^{4}x\cdot P(X=x)=\frac{10}{3},$$

$$\sigma_X^2=\mathbb{E}[(X-\mu_X)^2]=\sum_{x=2}^{4}(x-\mu_X)^2\cdot P(X=x)=\frac{44}{3},$$

We can also use $Var[X]=\mathbb{E}[X^2]-\mathbb{E}[X]^2$ to calculate the variance:

$$\mathbb{E}[X^2]=\sum_{x=2}^{4}x^2\cdot P(X=x)=\frac{35}{3}$$

$$\sigma_X^2=\mathbb{E}[X^2]-\mathbb{E}[X]^2=\frac{44}{3}.$$

We can find the distribution of Y first: range$(Y)=\{12,14,16\}$, and then we have

$$P(Y=y)=\begin{cases}\frac{1}{6}, \text{ if } y=12\\ \frac{1}{3}, \text{ if } y=14\\ \frac{1}{2}, \text{ if } y=16\end{cases}$$

$$\mu_Y=\mathbb{E}[Y]=\sum_{y=12,14,16}y\cdot P(Y=y)=\frac{44}{3},$$

$$\sigma_Y^2=\mathbb{E}[(Y-\mu_Y)^2]=\sum_{y=12,14,16}(y-\frac{44}{3})^2\cdot P(Y=y)=\frac{20}{9}$$

or using the other formula for the variance:

$$\mathbb{E}[Y^2]=\sum_{y=12,14,16}y^2\cdot P(Y=y)=\frac{652}{3}$$

$$\sigma_Y^2=\mathbb{E}[Y^2]-\mathbb{E}[Y]^2=\frac{20}{9}.$$

or using $Y=2X+8$:

$$\mathbb{E}[Y]=[aX+b]=a\mathbb{E}[X]+b=\frac{44}{3}$$

$$\sigma_Y^2=Var[aX+b]=a^2Var[X]=\frac{20}{9}.$$

Remark 3.4.1. It's not meaningful to compare $Var[X]$, and $Var[Y]$.

Remark 3.4.2. What does $\mu_X = \frac{10}{3}$ mean? Suppose we repeat the experiment a lot of times, we will have many 2's, 3's, and 4's but with different number of occurrences. We will have $\frac{10}{3}$ if we take the average.

Remark 3.4.3. From the previous example, we have $\mathbb{E}[X^2] \neq E[X]^2$, but what if $\mathbb{E}[X^2] = \mathbb{E}[X]^2$? Then $\text{Var}(X) = 0$, meaning that there's no variability around the mean, i.e., all points concentrate on the mean point. More rigorously,

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

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

3.5 I.I.D. Random Variables

Definition 3.5.1 (*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 3.5.1. The above three statements are equivalent. It mimics the definition of independent events.

Definition 3.5.2 (*Identical Distribution*)**.** X and Y are called *identically distributed* if

$$P_X = P_Y.$$

Remark 3.5.2. When we say X and Y are *I.I.D.*, we mean they are independent and identically distributed.

Theorem 3.5.1 (*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

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

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

Example 3.5.1. Let X and Y be i.i.d. random variables with the following distribution:

$$P(X = x) = \begin{cases} \frac{1}{6}, \text{ if } x = 2 \\ \frac{1}{3}, \text{ if } x = 3 \\ \frac{1}{2}, \text{ if } x = 4 \end{cases}$$

Define $Z = X + Y$. Given $\mathbb{E}[X] = \frac{10}{3}$, and $Var[X] = \frac{5}{9}$, find $\mathbb{E}[Z]$ and $Var[Z]$.

$\mathbb{E}[Z] = 2\mathbb{E}[X] = \frac{20}{3}$, and $Var[Z] = 2Var[X] = \frac{10}{9}$. Alternatively, we first find the distribution for Z: range$(Z) = \{4, 5, 6, 7, 8\}$

$$P(Z = 4) = P(X + Y = 4) = P(X = 2, Y = 2) = P(X = 2) \cdot P(Y = 2) = (\frac{1}{6})^2$$
$$P(Z = 5) = P(X + Y = 5) = P(X = 2, Y = 3) + P(X = 3, Y = 2) = 2 \cdot \frac{1}{6} \cdot \frac{1}{3}$$
$$P(Z = 6) = P(X = 2, Y = 4) + P(X = 3, Y = 3) + P(X = 4, Y = 2)$$
$$P(Z = 7) = P(X = 3, Y = 4) + P(X = 4, Y = 3)$$
$$P(Z = 8) = P(X = 4, Y = 4)$$

And then calculate the expectation and variance using their definitions.

3.6 Bernoulli Random Variables

Definition 3.6.1 (*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 3.6.2 (*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 3.6.1 (*Probability Distribution 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 3.6.2 (*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 3.6.1. 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. We don't use definitions like 3.6.1 for the proof.

3.7 Binomial Random Variables

Definition 3.7.1 (*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 3.7.1 (*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 $x \in$ 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 3.7.1. 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 3.7.2 (*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. □

Example 3.7.1. There're 3 independent parts in a machine, at least 2 parts must function s.t. the machine would work, and Y is the number of failed parts. Each part will fail with probability 0.1. At the end, the total cost for repairing is given by $R = 4Y + 5$. Find the expected value and variance for Y.

$$\begin{aligned} Y \sim Binomial(3, 0.1) &\Rightarrow \mathbb{E}[Y] = np = 3 \cdot 0.1 = 0.3 \\ &\Rightarrow Var[Y] = np(1-p) = 3 \cdot 0.1 \cdot 0.9 = 0.27 \end{aligned}$$

Find the expected value and variance for R.

$$\begin{aligned} &\mathbb{E}[R] = \mathbb{E}[4Y + 5] = 4\mathbb{E}[Y] + 5 = 4 \cdot 0.3 + 5 = 6.2 \\ &Var[R] = Var[4Y + 5] = 4^2 Var[Y] = 16 \cdot 0.27 = 4.32 \end{aligned}$$

Suppose the machine doesn't work, what's the probability that all parts failed?

$$\begin{aligned} P(Y = 3 | \text{“not working”}) &= \frac{P(Y = 3 \cap \text{“not working”})}{P(\text{“not working”})} \\ &= \frac{P(Y = 3 \cap (Y = 3 \cup Y = 2))}{P(Y = 3 \cup Y = 2)} \\ &= \frac{P(Y = 3)}{P(Y = 2) + P(Y = 3)} \\ &= \frac{\binom{3}{3} 0.1^3}{\binom{3}{2} 0.1^2 \cdot 0.9 + \binom{3}{3} 0.1^3} \end{aligned}$$

Suppose $R = 2Y^2 + 5$, what's the expected value and variance of R?

$$\begin{aligned} \mathbb{E}[R] \quad &= \mathbb{E}[2Y^2 + 5] \\ &= 2\mathbb{E}[Y^2] + 5, \text{ where } \mathbb{E}[Y^2] = Var[Y] + \mathbb{E}[Y]^2 \end{aligned}$$

$$\begin{aligned} Var[R] &= Var[2Y^2 + 5] \\ &= 4Var[Y^2] \\ &= 4(\mathbb{E}[Y^4] - \mathbb{E}[Y^2]^2), \text{ where } \mathbb{E}[Y^4] \text{ is calculated by definition} \end{aligned}$$

Remark 3.7.1. Note that, normally we will never calculate variance using its definition.

3.8 Geometric Random Variables

Definition 3.8.1 (*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 3.8.1 (*Probability Distribution 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 3.8.2 (*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. □

Example 3.8.1. The cost of an experiment is \$1000, and if an experiment fails, extra \$300 will be required. Suppose the probability for a successful trail is 0.2, assuming all trails are independent. Trails continue until a success. What's the probability that it takes at least 3 trails until the first success? At least 10 trails?

Let X be the number of trials until the first success, $X \sim Geometric(0.2)$.

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - \sum_{x=1}^{2} P(X = x) = 1 - \sum_{x=1}^{2} 0.8^{x-1} \cdot 0.2$$
$$P(X \geq 10) = 1 - P(X \leq 9) = 1 - \sum_{x=1}^{9} P(X = x) = 1 - \sum_{x=1}^{9} 0.8^{x-1} \cdot 0.2$$

Or equivalently,

$P(X \geq 3) = 0.8^2$, since $X \geq 3$ means all failures for the first two trails

$P(X \geq 10) = 0.8^9$

What's the expected value and variance of the cost until a successful trail?

Let Y be the cost until a successful trail, we have $Y = 1300(X - 1) + 1000 = 1300X - 300$,

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[1300X - 300] = 1300\mathbb{E}[X] - 300 = 1300 \cdot \frac{1}{0.2} - 300 \\ Var[Y] &= Var[1300X - 300] = 1300^2 Var[X] = 1300^2 = 1300^2 \cdot \frac{1 - 0.2}{0.2^2} \end{aligned}$$

Given that the first 9 trails are all failures, what's the probability that it will take at least 3 more trails until the first success?

$$\begin{aligned} P(X \geq 12 | X \geq 10) &= \frac{P(X \geq 12)}{P(X \geq 10)} \\ &= \frac{(1 - 0.2)^{11}}{(1 - 0.2)^9} \\ &= P(X \geq 3) \end{aligned}$$

Lemma 3.8.3. Let $X \sim \mathit{Geometric}(p)$, 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 3.8.1. $P(X \geq x)$ measures the probability that the first success doesn't occur on the first $x - 1$ trials.

Lemma 3.8.4 (*Memoryless Property*)**.** Let $X \sim \mathit{Geometric}(p)$, 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 3.8.2. $P(X > a+b|X > a)$ 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 $P(Y > b)$.

3.9 Negative Binomial Random Variable

Definition 3.9.1 (*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 $X \sim NegativeBinomial(r, p)$.

Lemma 3.9.1 (*Probability Distribution 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 3.9.1. 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 3.9.2 (*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

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

□

Example 3.9.1. A company has 3 jobs to fill, each applicant has 0.2 probability to be chosen, assuming they're independent. What's the probability that the 10^{th} applicant gets the last job?

Let $X \sim NegBin(3, 0.2)$ be the number of trials until the 3^{rd} success, then

$$P(X = 10) = \binom{9}{2} 0.2^3 \cdot 0.8^7$$

What's the probability that the 10th applicant gets a job?
Let $X_i \sim NegBin(i, 0.2)$ be the number of trials until the i^{th} success, for $i = 1, 2, 3$, then

$$P(X_1 = 10) + P(X_2 = 10) + P(X_3 = 10) = 0.8^9 \cdot 0.2 + \binom{9}{1} 0.8^8 \cdot 0.2^2 + \binom{9}{2} 0.8^7 \cdot 0.2^3$$

3.10 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, $X \sim \mathit{Binomial}(n,p)$, 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 $\mathit{Poisson}(\lambda)$ random variable, where $\mu_X = \lambda$.

Definition 3.10.1 (*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 $\mathit{Poisson}(\lambda)$ random variable.

Lemma 3.10.1 (*Probability Distribution of Poisson RVs*)**.** Let $X \sim \mathit{Poisson}(\lambda)$, then

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

Lemma 3.10.2 (*Expectations and Variance of Poisson RVs*)**.** Let $X \sim \mathit{Poisson}(\lambda)$, then

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

Proof. Skipped for now. □

Theorem 3.10.1. Consider n independent $\{X_i \sim \mathit{Poisson}(\lambda_i)\}_{i=1}^n$ random variables, then

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

Proof. Skipped for now. □

Example 3.10.1. Customers arrive at a place according to a Poisson distribution at an average of seven per hour. During a given hour, what's the probability that no more than 2 customers arrive?

$$P(X \leq 2) = \sum_{x=0}^2 P(X=x) = \sum_{x=0}^2 \frac{7^x}{x!} e^{-7}$$

Find the probability that exactly 2 customers arrive between $2-4$ pm.

Let X_i be the number of customers arrive during the i^{th} hour, i.e. $X_i \sim Poisson(7)$. Assume X_1 and X_2 are independent.

$$\begin{aligned} P(X_1 + X_2 = 2) &= P(X_1 = 0, X_2 = 2) + P(X_1 = 1, X_2 = 1) + P(X_1 = 2, X_2 = 0) \\ &= P(X_1 = 0) \cdot P(X_2 = 2) + P(X_1 = 1) \cdot P(X_2 = 1) + P(X_1 = 2) \cdot P(X_2 = 0) \end{aligned}$$

Or equivalently, $X = X_1 + X_2 \sim Poisson(14)$, then

$$P(X = 2) = \frac{14^2}{2!} e^{-14}.$$

Theorem 3.10.2 (*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. □

Example 3.10.2. 1000 independent trails, and a success comes with probability 0.005 for each trail. What's the probability of at least 4 successes? Let $X \sim Binomial(1000, 0.005)$ be the number of successes.

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^{3} \binom{1000}{x} 0.005^x \cdot 0.995^{1000-x},$$

or approximately, $X \sim Poisson(1000 \cdot 0.005 = 5)$, then

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^{3} \frac{5^x}{x!} e^{-5}$$

Theorem 3.10.3. 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 3.10.4 (*Poisson Counting Process*)**.** Let $X \sim Poisson(\lambda)$ 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

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

assuming the successes occur uniformly over the time interval.

Proof. Skipped for now. □

3.11 Hypergeometric Random Variables

Definition 3.11.1 (*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 $Hypergeometric(N, r, n)$ random variable.

Lemma 3.11.1 (*Probability Distribution of Hypergeometric RVs*)**.** Let $X \sim Hypergeometric(N, r, n)$, then

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

Proof. Skipped for now. □

Theorem 3.11.1 (*Expectations and Variance of Hypergeometric RVs*)**.** Let $X \sim Hypergeometric(N, r, n)$, 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. □

4 Continuous Random Variables

4.1 Cumulative Distribution Function

Definition 4.1.1 (*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 4.1.1 (*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 4.1.2 (*Probability Masses Calculation*)**.** Let X be any random variable, then its *probability mass* at a can be calculated using its cdf as follows:

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

Example 4.1.1. Consider $X \sim Binomial(2, \frac{1}{2})$, find the cdf of X.
Then we have $P(X = x) = \binom{2}{x}(\frac{1}{2})^2$, specifically,

$$\begin{aligned} P(X = 0) &= \frac{1}{4} \\ P(X = 1) &= \frac{1}{2} \\ P(X = 2) &= \frac{1}{4} \end{aligned}$$

and hence,

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{1}{4}, & \text{if } 0 \leq x < 1 \\ \frac{3}{4}, & \text{if } 1 \leq x < 2 \\ 1, & \text{if } x \geq 2 \end{cases}$$

Note that if the RV is discrete, the cdf will be a step function, where the steps are the corresponding probability masses.

Example 4.1.2. Consider a RV X with cdf as follows:

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ x^2, & \text{if } 0 \leq x < y \\ 1, & \text{if } y \geq 1 \end{cases}$$

Verify this is a valid cdf.

- $F_X(-\infty) = 0$
- $F_X(\infty) = 1$
- Non-decreasing
- Fully continuous, and hence right continuous

Therefore, it's a valid cdf. Note that there're no steps, no probability masses, i.e., $P(X = x) = 0$ for all x.

4.2 Continuous Random Variables

Definition 4.2.1 (*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 4.2.2 (*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,

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

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

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

□

Lemma 4.2.1 (*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.
- $\int\limits_{-\infty}^{\infty} f_X(x)dx = \lim_{x\to\infty} F_X(x) = 1$
- $P(X = a) = F_X(a) - \lim_{x\to a^-} F_X(x) = 0$, 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

□

Definition 4.2.3 (*Percentiles*)**.** Let X be a continuous random variable, then the $100p^{th}$ percentile 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.

4.3 Expectations

Definition 4.3.1 (*Raw Moments*)**.** Let X be a continuous random variable, then the k^{th} *raw moment* of X is defined as

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

Definition 4.3.2 (*Expectations*)**.** Let X be a continuous random variable, then the *expected value* of X is defined as its first raw moment, i.e.,

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

Theorem 4.3.1 (*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 4.3.2 (*Scaling & Shifting on Expectations*)**.** Let a, b be constants, we have the following:

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

Theorem 4.3.3 (*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.

Remark 4.3.1. Other theorems under the context of discrete RVs remains the same, which we will ignore here.

4.4 Variance

Definition 4.4.1 (*Central Moments*)**.** Let X be a continuous random variable, then the k^{th} *central moment* of X is defined as

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

Definition 4.4.2 (*Variance*)**.** Let X be a continuous random variable, then the *variance* of X is defined as its second raw moment, i.e.,

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

Theorem 4.4.1 (*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 4.4.2 (*Linear Combinations of RVs*)**.** Let $\{X_i\}_{i=1}^n$ be independent random variables. If $Y = a_1 X_1 + ... + a_n X_n + b$, 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

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

Remark 4.4.1. Other theorems under the context of discrete RVs remains the same, which we will ignore here.