MATH3280: Introductory Probability

nablamath

April 30, 2022

Remarks

(1) Context of this document is based on university course *MATH3280: Introductory Probability* from *Department of Mathematics, The Chinese University of Hong Kong*. The original source can be found at `https://www.math.cuhk.edu.hk/course`. The author does not own the source.

(2) This document is assumed unavailable for unauthorized parties that have not attended the university course. It is prohibited to share, including distributing or copying this document to unauthorized parties in any means for any non-academic purpose.

(3) Context of this doucment may not be completely accurate. The author assumes no responsibility or liability for any errors or omissions in the context of this document.

Source

The latest version of this document can be found at `https://www.github.com/nablamath/notes`.

Contents

1 Axioms of Probability

1.1 Introduction to Probability

1.1.1 Definition of Probability

Definition 1.1. **Probability** is a study of random behaviours in mathematics.

Probability has a history of more than 300 years, which comes from gambling and games of chance.

1.1.2 Terminologies of Probability

Probability involves **random experiments** and **outcomes**.

Example 1.2. Below are examples of random experiments and their corresponding outcomes:

(a) Toss a coin once to obtain a head or a tail.

(b) Roll a die once to see the number of the top face.

(c) Randomly choose a student in the class and measure the height of the student.

Definition 1.3. **Sample space**, usually denoted by S, is the set of all possible outcomes of a random experiment.

Example 1.4. Below are examples of sample spaces of random experiments:

(a) For tossing a coin once, the sample space $S = \{H, T\}$ where H represents head and T represents tail.

(b) For tossing a coin twice, the sample space $S = \{HH, HT, TH, TT\}$.

(c) For rolling a die 4 times to record the numbers appearing in the top face, the sample space $S = \{(x_1, x_2, x_3, x_4) \mid x_i \in \{1, 2, 3, 4, 5, 6\}, 1 \leq i \leq 4\}$.

(d) For the height of a random student in the campus (in meters), $S = \{0 \leq x \leq \infty\}$.

Definition 1.5. Let S denote the sample space of a random experiment, then any subset E of S is called an **event** of the experiment.

If the outcome of an experiment is contained in event E, then E is said to be **occured**. Note that some special events always occurs or not occurs. For example, a **null event**, denoted by ϕ, never occurs in any random experiment.

1.2 Probability Operations

1.2.1 Basic Operations on Events

Below are some basic operations on events:

Proposition 1.6. Let E and F are events of a random experiment, then the following applies:

(a) **Intersection** of E and F, denoted by $E \cap F$.

(b) **Union** of E and F, denoted by $E \cup F$.

(c) **Complement** of E, denoted by E^c.

In order to understand basic operations on events, **Venn diagram** is introduced. In Venn diagram, a rectangular region represents a sample space, while a (usually circular) region within the rectangular region represents an event.

1.2.2 Laws of Event Operations

There are three basic laws of event operations:

Proposition 1.7. The **Communtative Law** states that $E \cap F = F \cap E$ and $E \cup F = F \cup E$.

Proposition 1.8. The **Associative Law** states that $E \cup (F \cup G) = (E \cup F) \cup G$ and $E \cap (F \cap G) = (E \cap F) \cap G$.

Proposition 1.9. The **Distributive Law** states that $E \cap (F \cup G) = (E \cap F) \cup (E \cup G)$ and $E \cup (F \cap G) = (E \cup F) \cap (E \cup G)$.

Besides the three basic laws, another common law is the **De Morgan's Laws**.

Theorem 1.10. De Morgan's Laws state that

$$\left(\bigcup_{n=1}^{\infty} E_n \right)^c = \bigcap_{n=1}^{\infty} E_n^c$$

$$\left(\bigcap_{n=1}^{\infty} E_n \right)^c = \bigcup_{n=1}^{\infty} E_n^c$$

Proof. Note that

$$\begin{aligned} x \in \bigcup_{n=1}^{\infty} E_n{}^c &\Leftrightarrow x \in S, x \notin \bigcup_{n=1}^{\infty} E_n \\ &\Leftrightarrow x \in S, x \in E_n \text{ for all } n \\ &\Leftrightarrow x \in S \setminus E_n \text{ for all } n \\ &\Leftrightarrow x \in \bigcap_{n=1}^{\infty} E_n^c \end{aligned}$$

The other law can be proved by using the same technique.

1.3 Axioms and Properties of Probability

1.3.1 Axiomatic Approach to Probability

It is an important question knowing how to define the probability of an event, or how likely an event will happen in an experiment. An inituitive approach in define the probability is to repeat the experiment n times, so if the experiment is done for numerous times and $n(E)$ is number of times that an event E has occured, then the probability of E

$$P(E) = \lim_{n\to\infty} \frac{n(E)}{n}$$

This is a natural idea to define probability as the probability should be higher if the chance that the event will happen is higher. However, one have to consider the drawbacks of the definition above. It is unsure that the limit exists or not, and even if the limit exists, it is dependent of the experiments conducted.

The definition above may not be vigorous enough, but it is understandable. However, in order to have a vigorous definition, **axiomatic approach to probability** is used. Such approach is proposed by Kolmogrov in the 20th century.

Definition 1.11. Let S be the sample space of an experiment, then probability P on S is a function which assigns a value to each event E of S such that the following axioms hold:

(a) **Axiom** 1

$$0 \leq P(E) \leq 1 \quad \text{for all } E \subset S$$

(b) **Axiom** 2

$$P(S) = 1$$

(c) **Axiom** 3

Let $E_1, E_2, \cdots$ be a sequence of events that are mutually exclusive, or in other words, $E_i \cap E_j = \phi \quad \text{if } i \neq j$, then

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n)$$

Note that the last axiom of the definition above is also called the **countable additivity of probability**.

1.3.2 Properties of Probability

Proposition 1.12. A null event has probability of 0, or $P(\phi) = 0$.

Proof. Let $(E_n)_{n=1}^{\infty}$ be a sequence of event by $E_1 = S$ and $E_2 = E_3 = \cdots = \phi$. Since the events are mutually disjoint, then by Axiom 3,

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n) = P(S) + P(\phi) + P(\phi) + \cdots$$

Note that left hand side of the equation is 1 by Axiom 1, and $P(S) = 1$ by Axiom 2. Therefore $P(\phi) = 0$ in order to hold the equation above.

Proposition 1.13. Let $E \subset S$ be an event, then probability of complement of E, $P(E^c) = 1 - P(E)$.

Proof. Let $(E_n)_{n=1}^{\infty}$ be a sequence of event by $E_1 = E$, $E_2 = E^c$ and $E_3 = E_4 = \cdots = \phi$. Since the events are mutually disjoint, then by Axiom 3,

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n) = P(E) + P(E^c) + P(\phi) + P(\phi) + \cdots$$

Note that left hand side of the equation is 1 by Axiom 1, and $P(S) = 1$ by Axiom 2. Therefore $P(E) + P(E^c) = 1$ in order to hold the equation above. Rearrange the equation finishes the proof.

The following proposition is the **finite additivity** of disjoint events:

Proposition 1.14. Let $E_1, E_2, \cdots, E_n$ be disjoint events, then

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

Proof. Let $E_{n+1} = E_{n+2} = \cdots = \phi$, then $(E_n)_{n=1}^{\infty}$ is disjoint. By Axiom 3,

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

and the equation is reduced to desired equation by using Axiom 1 and 2.

Proposition 1.15. Let E and F be two events, then $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Proof. Venn diagram shows the result explicitly. On the other hand, note that $E \cup F = E \cup (F \setminus E)$. By Axiom 3 and *Proposition 1.14*,

$$P(E \cup F) = P(E \cup (F \setminus E)) = P(E) + P(F \setminus E)$$

Also notice that $F = (F \setminus E) \cup (E \cap F)$, hence

$$P(F) = P(F \setminus E) + P(E \cap F)$$

Combine both equations by substituting $P(F \setminus E)$ gives the result.

Proposition 1.16. Let E and F be two events. If $E \subset F$, then $P(E) \leq P(F)$.

Proof. Note that $F = E \cup (F \setminus E)$, then

$$P(F) = P(E) + P(F \setminus E) \geq P(E)$$

The following proposition, which is also called the **inclusion-exclusion identity**, is the generalization of *Proposition 1.15*:

Proposition 1.17. Let $E_1, E_2, \cdots, E_n$ be events, then

$$\begin{aligned} P(E_1 \cup E_2 \cup \cdots \cup E_n) = &\sum_{i=1}^{n} P(E_i) - \sum_{i_1<i_2} P(E_{i_1} \cap E_{i_2}) \\ &- \sum_{i_1<i_2<i_3} P(E_{i_1} \cap E_{i_2} \cap E_{i_3}) \\ &+ \cdots + (-1)^{n+1} P(E_1 \cap E_2 \cap \cdots \cap E_n) \\ = &\sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1<i_2<\cdots<i_r} P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_r}) \end{aligned}$$

Example 1.18. Let $E, F \in S$ be events. Suppose $P(E) = 0.8$ and $P(F) = 0.9$, prove that $P(E \cap F) \geq 0.7$.

Answer. By the inclusion-exclusion identity,

$$\begin{aligned} P(E \cap F) &= P(E) + P(F) - P(E \cup F) \\ &\geq P(E) + P(F) - 1 = 0.7 \end{aligned}$$

Example 1.19. Let $E, F \in S$ be events. Suppose $P(E) = 0.8$, $P(F) = 0.9$ and $P(E \cap F) = 0.75$. Find the probability that exactly one of E and F has occured.

Answer. Let G be the required event. Note that $G = (E \setminus F) \cup (F \setminus E)$ is a union of mutually exclusive events, then

$$\begin{aligned} P(G) &= P(E \setminus F) + P(F \setminus E) \\ &= P(E) - P(E \cap F) + P(F) - P(F \cap E) = 0.2 \end{aligned}$$

1.3.3 Mutually Disjoint Events

Definition 1.20. Let $E_1, E_2, \cdots, E_n$ be events, then union of $(E_i)_{i=1}^n$ can be expressed as a union of mutually disjoint events $(F_i)_{i=1}^n$ by

$$\begin{cases} F_1 = E_1 \\ F_k = E_k \setminus \bigcup_{i=1}^{k-1} E_i \end{cases}$$

Note that $F_n \subset E_n$ for any n. The above definition helps to convert any set of events into a set of mutually exclusive events, so that axioms and properties of probability can be used.

Proposition 1.21. Let $(F_n)_{n=1}^\infty$ as in *Definition 1.20*, then $F_n \cap F_m = \phi$ for any $n \neq m$.

Proof. Without loss of generality, assume $n > m$, then

$$\begin{aligned} F_n &= E_n \setminus (E_1 \cup \cdots \cup E_m \cup \cdots \cup E_{n-1}) \\ F_n \cap E_m &= \phi \end{aligned}$$

but since $F_m \subset E_m$, $F_n \cap F_m = \phi$.

The proposition above ensures that $(F_n)_{n=1}^\infty$ are mutually exclusive.

Proposition 1.22. Let $(F_n)_{n=1}^{\infty}$ as in *Definition 1.20*, then

$$\bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i$$

The next proposition is the **countable subadditivity** of probability:

Proposition 1.23. Let $E_1, E_2, \cdots, E_n$ be events, then

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} P(E_n)$$

Proof. Let $(F_n)_{n=1}^{\infty}$ as in *Definition 1.20*, then by *Proposition 1.21*, *Proposition 1.22* and applying Axiom 3 to (F_n),

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = P\left(\bigcup_{n=1}^{\infty} F_n\right) = \sum_{n=1}^{\infty} P(F_n) \leq \sum_{n=1}^{\infty} P(E_n)$$

1.3.4 Continuity of Probability

Continuity of probability consists of two parts, which involves increasing and decreasing subsets of events respectively.

Proposition 1.24. Let $E_1 \subset E_2 \subset \cdots \subset E_n \subset \cdots$ be increasing, then

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} P(E_n)$$

Proof. Let $(F_n)_{n=1}^{\infty}$ as defined in *Definition 1.24* 1.22. Since $(F_n)_{n=1}^{\infty}$ are mutually disjoint,

$$\bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i = E_n$$

and

$$\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} E_i$$

Applying Axiom 3 to (F_n) gives

$$\begin{aligned} P\left(\bigcup_{n=1}^{\infty} E_n\right) &= P\left(\bigcup_{n=1}^{\infty} F_n\right) \\ &= \sum_{n=1}^{\infty} P(F_n) \\ &= \lim_{n\to\infty}\left(\sum_{i=1}^{n} P(F_i)\right) \\ &= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} F_i\right) \\ &= \lim_{n\to\infty} P(E_n) \end{aligned}$$

Proposition 1.25. Let $E_1 \supset E_2 \supset \cdots \supset E_n \supset \cdots$ be increasing, then

$$P\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} P(E_n)$$

Proof. Consider

$$1 - P\left(\bigcap_{n=1}^{\infty} E_n\right) = P\left(\left(\bigcap_{n=1}^{\infty} E_n\right)^c\right) = P\left(\bigcup_{n=1}^{\infty} E_n^c\right)$$

by De Morgan's Law. Apply the proof in *Proposition 1.24* gives

$$1 - P\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty}(1 - P(E_n))$$

Rearrange the equation finishes the proof.

1.3.5 Cardinality of Events

Definition 1.26. Let E be an event in S. The **cardinality** of event E, denoted by $\#E$, is the number of outcomes occured in the sample space of the experiment.

In many experiments, it can be assumed that all outcomes of the experiment have the same chance to occur, then $P(E) = \#E/\#S$.

Example 1.27. Find the probability that the sum of top faces of two dice is equal to 5.

Answer. Let E be the required event. Note that

$$S = \{(i, j) \mid i, j \in \{1, 2, 3, 4, 5, 6\}\}$$

and

$$E = \{(i,j) \in S \mid i+j=5\} = \{(1,4),(2,3),(3,2),(4,1)\}$$

Hence

$$P(E) = \frac{\#E}{\#S} = \frac{4}{36} = \frac{1}{9}$$

Example 1.28. A committee of 5 is to be selected from a group of 6 men and 9 women. If the selection is made randomly, find the probability that the committee consists of 3 men and 2 women.

Answer. Let E be the required event. Note that $P(E) = \#E/\#S$ where

$$\#S = \binom{15}{5},\ \#E = \binom{6}{3}\binom{9}{2}$$

Example 1.29. In the game of bridge, the entire deck of 52 cards is dealt out to 4 players. Find the probability that

(a) one of the players receives all 13 spades.

(b) each player receives an ace.

Answer. Part (a) Note that the cardinality of sample space is

$$\#S = \binom{52}{13}\binom{39}{13}\binom{26}{13}\binom{13}{13}$$

Let E be the required event, and E_i be the event where the i-th player receives all spades. Note that $E = \bigcup_{i=1}^{4} E_i$ with E_i being mutually exclusive. The i-th player first receives spades, other players share the remaining cards, then

$$\#E_i = \binom{13}{13}\binom{39}{13}\binom{26}{13}\binom{13}{13}$$

Therefore

$$P(E) = \frac{\#E}{\#S} = \frac{4\#E_i}{\#S}$$

Part (b) Let F be the required event. Each player first get an ace, then take 12 cards from the non-ace pile, then

$$\#F = \binom{4}{1}\binom{48}{12}\binom{3}{1}\binom{36}{12}\binom{2}{1}\binom{24}{12}\binom{13}{13}$$

and $P(F) = \#F/\#S$.

1.4 Conditional Probability and Independence

1.4.1 Conditional Probability

Below is an example of probability with given conditions.

Example 1.30. Two fair dice are rolled. Given that the first die is a 4, find the probability that the sum of two dice is 9.

Normally without additional conditions, the event that the sum of two dice is equal to 9, denoted by E, has outcomes $\{(3,6),(4,5),(5,4),(6,3)\}$. It is easy to find out $P(E) = 1/9$. Similarly, the event that the first die is a 4, denoted by F, has outcomes $\{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}$.

With conditions, it is assumed that F has occured. Note that $E \cap F = \{(4,5)\}$, so the probability of E given F is $1/6$.

Definition 1.31. Let E and F be two events in a random experiment. Suppose $P(F) > 0$, then the **conditional probability** of E given F, denoted by $P(E \mid F)$, is given by

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}$$

Note that if $P(F) = 0$, $P(E \mid F)$ is then not well-defined.

Example 1.32. A fair coin is flipped 3 times. Find the conditional probability that the third flip is a head, given that the first flip is a tail.

Answer. Let E be the event that the first lip is a tail and F be the event that the third flip is a head. By *Definition 1.31*,

$$P(F \mid E) = \frac{P(F \cap E)}{P(E)} = \frac{2/8}{4/8} = \frac{1}{2}$$

The following proposition is also called **multiplicative rule**.

Proposition 1.33. Let E_i be events, then

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2 \mid E_1)P(E_3 \mid (E_1 \cap E_2)) \cdots P(E_n \mid (E_1 \cap E_2 \cap \cdots \cap E_{n-1}))$$

Proof. Rearrange *Definition 1.31* gives

$$P(E_1 \cap E_2) = P(E_1)P(E_2 \mid E_1)$$

and can be extended to case of n events by induction.

1.4.2 Bayes Formula

Before introducing the main formula, consider the following proposition which is also known as **total probability formula**:

Proposition 1.34. Let E and F be two events, then

$$P(E) = P(F)P(E \mid F) + P(F^c)P(E \mid F^c)$$

Proof. Note that

$$E = (E \cap F) \cup (E \cap F^c)$$

and since both events in E are mutually exclusive,

$$P(E) = P(E \cap F) + P(E \cap F^c)$$

By multiplicative rule,

$$P(E) = P(F)P(E \mid F) + P(F^c)P(E \mid F^c)$$

This formula works for any event F, and it is said to be a conditioning method for calculating unconditional probability.

Definition 1.35. Let $E_1, E_2, \cdots, E_n$ be events, then the events are **exhausitive** if

$$\bigcup_{i=1}^{n} E_i = S$$

Below is a generalized formula of *Proposition 1.34*:

Corollary. Let E be an event, and $F_1, F_2, \cdots, F_n$ be mutually exclusive and exhausitive events, then

$$P(E) = \sum_{i=1}^{n} P(F_i)P(E \mid F_i)$$

Proof. Note that $E \cap F_i$ are mutually disjoint, so

$$\begin{aligned} P(E) &= \sum_{i=1}^{n} P(E \cap F_i) \\ &= \sum_{i=1}^{n} P(F_i)P(E \mid F_i) \end{aligned}$$

With the formula above, **Bayes formula** is introduced as below:

Theorem 1.36. Let $F_1, F_2, \cdots, F_n$ be mutually exclusive and exhausitive events, then

$$P(F_i \mid E) = \frac{P(F_i)P(E \mid F_i)}{\sum_{j=1}^{n} P(F_j)P(E \mid F_j)}$$

Proof. By total probability formula,

$$P(E) = \sum_{j=1}^{n} P(F_j)P(E \mid F_j)$$

The proof is finished by applying *Definition 1.31* with $P(E \cap F_i) = P(F_i)P(E \mid F_i)$.

Example 1.37. A bin contains 3 different types of disposable flashlights, which are type 1, 2 and 3. Each type of flashlight has a probability of 0.7, 0.4 and 0.3 respectively to give over 100 hours of use. Suppose 20% of the flashlights are type 1, 30% of the flashlights are type 2 and 50% of the flashlights are type 3.

(a) Find the probability that a randomly chosen flashlight will give more than 100 hours of use.

(b) Given that a flashlight lasted over 100 hours, what is the conditional probability that it was type 1, 2 or 3.

Answer. Part (a) Let E be the event that the flashlight gives more than 100 hours of use and F_j be the event that the flashlight is of type j. Note that $P(F_1) = 0.2$, $P(F_2) = 0.3$ and $P(F_3) = 0.5$, while $P(E \mid F_1) = 0.7$, $P(E \mid F_2) = 0.4$ and $P(E \mid F_3) = 0.3$, then

$$P(E) = \sum_{j=1}^{3} P(F_j)P(E \mid F_j) = 0.2(0.7) + 0.3(0.4) + 0.5(0.3) = 0.41$$

Part (b) Without loss of generality,

$$P(F_j \mid E) = \frac{P(F_j)P(E \mid F_j)}{P(E)}$$

by the definition of conditional probability.

Example 1.38. Two fair dice are rolled. Find the conditional probability that at least one of them is 6 given that the dice land on different numbers.

Answer. Let E be the event where at least one of the dice is 6, and F be the event that two dice land on different numbers. Note that $\#S = 36$, $\#F = 30$ and $\#(E \cap F) = 10$, so

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{10/36}{30/36} = \frac{1}{3}$$

1.4.3 Conditional Independence

For most of the cases, $P(E \mid F)$ is not equal to $P(E)$, but there are some special cases where they are equal.

Definition 1.39. Let E and F be events, then E is said to be **independent** of F if $P(E \mid F) = P(E)$.

The following proposition shows the symmetric property of conditional independence:

Proposition 1.40. Let E and F be events, then E and F are independent if $P(E \cap F) = P(E)P(F)$.

Proof. By the definition of conditional probability in *Definition 1.32*, if E is independent of F, then

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} = P(E)$$

which means $P(E \cap F) = P(E)P(F)$. Apply the definition again gives $P(F \mid E) = P(F)$, hence F is also independent of E.

Example 1.41. A card is randomly chosen from a deck to 52 playing cards. Let E be the event that the chosen card is an ace, and F be the event that the chosen card is a spade. Check whether E and F are independent.

Answer. Since

$$P(E \cap F) = \frac{1}{52} = \frac{1}{13}\left(\frac{1}{4}\right) = P(E)P(F)$$

E and F are independent.

Below are some properties of conditional independence:

Proposition 1.42. If E and F are independent events, then the following applies:

(a) E and F^c are independent.

(b) E^c and F^c are independent.

Proof. Since E and F are independent, $P(E \cap F) = P(E)P(F)$. Notice that

$$\begin{aligned} P(E \cap F^c) &= P(E) - P(E \cap F) \\ &= P(E) - P(E)P(F) \\ &= P(E)(1 - P(F)) = P(E)P(F^c) \end{aligned}$$

then E and F^c are independent. It then follows that E^c and F^c are independent by applying the same method on E.

It is also important to discuss about conditional independence for more than 2 events. For simplicity, definition of 3 events is first introduced:

Definition 1.43. Let E, F and G are events, then they are independent if the following are satisfied:

(a) $P(E \cap F \cap G) = P(E)P(F)P(G)$.

(b) For any two events, they are independent to each other. In other words, E and F are independent, E and G are independent, and F and G are independent.

Definition 1.44. Let $\{E_1, E_2, \cdots, E_n\}$ be a finite family of events, then they are independent if

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = \prod_{i=1}^{n} P(E_i)$$

and for any subfamily $\{E_{j_1}, E_{j_2}, \cdots, E_{j_k}\}$, they are independent.

Furthermore, an infinite family of events are said to be independent if any finite subfamily of events are independent.

Definition 1.45. A random experiment consists of **subexperiments** if the events $E_1, E_2, \cdots, E_n$ are independent where E_i is an event whose occurence only depends on the i-th subexperiment.

2 Random Variables

2.1 Introduction to Random Variables

2.1.1 Definition of Random Variables

Definition 2.1. In a random experiment, a **random variable**, denoted by X, is a real-valued function defined on the sample space S.

With the definition above, random variable X is a function that maps from the sample space S to the real number set $\mathbb{R}$. Since S has outcomes of random phenomenon, X that depends on S is random. The following example demonstrates its randomness:

Example 2.2. Three fair coins are flipped. Let X be the number of heads appeared. Note that $X = 2$ if the outcome is (H, T, H), and $X = 0$ if the outcome is (T, T, T).

Random variable X does not always reflect the outcome explicitly, as shown in the following example:

Example 2.3. Two fair dice are rolled. Let X be the product of the two numbers appeared. Note that $X = 12$ if the outcome is $(2, 6)$ or $(4, 3)$.

2.1.2 Discrete Random Variables

Definition 2.4. Let X be a random variable, then X is said to be **discrete** if it takes at most countably many different values.

With a discrete random variable, it is also important to know how to measure the probability of values of X. There is a function that can fulfill the purpose above.

Definition 2.5. Let X be a discrete random variable. For any $a \in \mathbb{R}$, the **probability mass function** of a, denoted by $p(a)$, is defined as

$$p(a) = P(\{X = a\}) = P(\{\omega \in S \mid X(\omega) = a\})$$

In general, p is called the probability mass function of X.

Proposition 2.6. Let $x_1, x_2, \cdots, x_n$ be all possible values of a discrete random variable X, then

$$p(a) = 0 \quad \text{if } a \notin \{x_1, x_2, \cdots, x_n\}$$

and

$$\sum_{i=1}^{n} p(x_i) = 1$$

Proof. Let $E_i = \{\omega \in S \mid X(\omega) = x_i\}$, then E_i are mutually exclusive. Moreover, since $\bigcup_{i=1}^{n} E_i = S$,

$$1 = P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) = \sum_{i=1}^{n} p(x_i)$$

2.1.3 Expected Value of Discrete Random Variables

Definition 2.7. Let X be a discrete random variable, and p be the probability mass function of X, then the **expected value** of X, denoted by $E[X]$, is defined as

$$E[X] = \sum_{p(a)>0} ap(a) = \sum_{i} x_i p(x_i)$$

where x_i are possible values of X.

From the definition above, it can be seen that expected value of X is a weighted average of X. The weight depends on the probability of occurence of each value. Therefore, expected value of X is sometimes called the mean of X.

Example 2.8. Let X be a discrete random variable to represent number of heads appeared in three fair coins, then the expected value of X

$$E[X] = 0\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) = \frac{3}{2}$$

Now let X be a discrete random variable on S, $g : \mathbb{R} \to \mathbb{R}$ be a function such that $Y = g(X)$, where Y is a function on S. Observe that if x_i are possible values of X, then $g(x_i)$ are possible values of Y, so Y is another discrete random variable on S. The following proposition shows a way to calculate expected value of Y simply by values in X:

Proposition 2.9. Let X and Y be discrete random variables on S where $Y = g(X)$ by a function $g : \mathbb{R} \to \mathbb{R}$, then

$$E[Y] = \sum_{i} g(x_i)p(x_i)$$

Proof. Let y_i be possible unique values of Y. Grouping $g(x_i)$ with the same value gives

$$\begin{aligned}\sum_i g(x_i)p(x_i) &= \sum_j \sum_{g(x_i)=y_j} g(x_i)p(x_i) \\ &= \sum_j \sum_{g(x_i)=y_j} y_j p(x_i) \\ &= \sum_j y_j \sum_{g(x_i)=y_j} p(x_i) \\ &= \sum_j y_j \sum_{g(x_i)=y_j} P(\{X = x_i\}) \\ &= \sum_j y_j P(\{Y = y_j\}) \\ &= E[Y]\end{aligned}$$

Corollary. Let X be a discrete random variable and $a, b \in \mathbb{R}$, then

$$E[aX + b] = aE[X] + b$$

Proof. Let $g(x) = ax + b$ be a function. By *Proposition 2.9*,

$$E[aX + b] = E[g(X)] = \sum_i g(x_i)p(x_i)$$

where x_i are possible different values of X. Expand g gives

$$\begin{aligned}\sum_i g(x_i)p(x_i) &= \sum_i (ax_i + b)p(x_i) \\ &= a\sum_i x_i p(x_i) + b\sum_i p(x_i) \\ &= aE[X] + b\end{aligned}$$

2.1.4 Variance of Discrete Random Variables

Definition 2.10. Let X be a discrete random variable, then the **variance** of X, denoted by $\mathrm{Var}(X)$, is defined as

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

where $\mu = E[X]$.

Note that variance of X is sometimes written as $V(X)$ for simplicity. Variance of X describes how X is spread out from its mean value μ.

Proposition 2.11. Let X be a discrete random variable, then $\text{Var}(X) = E[X^2] - \mu^2$

Proof. Note that

$$\begin{aligned}\text{Var}(X) &= E[(X-\mu)^2] \\ &= \sum_i (x_i - \mu)^2 p(x_i) \\ &= \sum_i (x_i^2 - 2\mu x_i + \mu^2) p(x_i) \\ &= \sum_i x_i^2 p(x_i) - 2\mu \sum_i x_i p(x_i) + \mu^2 \sum_i p(x_i) \\ &= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2\end{aligned}$$

Corollary. Let X be a discrete random variable, then $E[X^2] \geq (E[X])^2$.

Proof. From the definition of variance, $\text{Var}(X) \geq 0$, then $E[X^2] - \mu^2 \geq 0$ implies the result.

2.2 Common Types of Discrete Random Variables

2.2.1 Bernoulli Random Variables

Consider a random experiment where the outcomes can be classified by either a success or a failure, then the following definition applies:

Definition 2.12. Let X be a discrete random variable where

$$X = \begin{cases} 1 & \text{if the outcome is a success} \\ 0 & \text{if the outcome is a failure} \end{cases}$$

then X is called a **Bernoulli random variable** with parameter $p = P(\{X = 1\})$.

Note that $p(0) + p(1) = 1$ and $p(a) = 0$ if a is neither 0 nor 1. Since the expected values $E[X] = p$ and $E[X^2] = p$, then the variance $V(X) = E[X^2] - (E[X])^2 = p - p^2$.

2.2.2 Binomial Random Variables

Consider a random experiment with n subexperiments and each subexperiment results in either a success or a failure, then the following definition applies:

Definition 2.13. Let X be a discrete random variable that is equal to the number of successes in a random experiment with n subexperiments, then X is called a **Binomial random variable** with parameters $(n, p = P(\text{success}))$.

Consider the following example:

Example 2.14. For $n = 2$, the possible outcomes of the experiments are (S, S), (S, F), (F, S) and (F, F) where S indicates a success and F indicates a failure. If the probability of a success in each subexperiment is p,

$$\begin{cases} P(\{X = 0\}) = P(\{(F, F)\}) = (1 - p)^2 \\ P(\{X = 1\}) = P(\{(S, F), (F, S)\}) = 2p(1 - p) \\ P(\{X = 2\}) = P(\{(S, S)\}) = p^2 \end{cases}$$

Proposition 2.15. Let X be a binomial random variable with parameters (n, p), then

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

for $k = 0, 1, \cdots, n$.

Proof. Note that for any sequence of outcomes of n subexperiments in which k of them results in successes, there will be $n-k$ failures, so each sequence has probability $p^k(1 - p)^{n-k}$. By combinatorics, there are C_k^n sequences that have k successes. The proof is finished by multiplying the results.

Therefore the name of Binomial random variable comes from binomial constant which is also applied to Binomial theorem

$$(x + y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

Proposition 2.16. Let X be a binomial random variable with parameters (n, p), then for $k \geq 1$,

$$E[X^k] = npE[(Y + 1)^{k-1}]$$

where Y is another binomial random variable with parameters $(n - 1, p)$.

Proof. By *Definition 2.13*,

$$\begin{aligned}
E[X^k] &= \sum_{i=0}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} \\
&= \sum_{i=1}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} \\
&= \sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^i (1-p)^{n-i} \\
&= np \sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} \\
&= np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} \\
&= npE[(Y+1)^{k-1}]
\end{aligned}$$

Corollary. Let X be a binomial random variable with parameters (n, p), then $E[X] = np$, $E[X^2] = np((n-1)p+1)$ and $\mathrm{Var}(X) = n(p-p^2)$.

Proof. Note that

$$E[X] = npE[(Y+1)^0] = np$$

and

$$E[X^2] = npE[Y+1] = np(E[Y]+1) = np(np-p+1)$$

Therefore

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = np(np-p+1) - (np)^2 = n(p-p^2)$$

A binomial random variable X with parameter (n, p) can be expressed as

$$X = X_1 + X_2 + \cdots + X_n$$

where X_i are independent Bernoulli random variables.

2.2.3 Poisson Random Variables

Definition 2.17. Let $\lambda > 0$, then a **Poisson random variable** with parameter λ is a random variable X that takes nonnegative integers $i = \{0, 1, \cdots\}$ such that

$$P(\{X = i\}) = e^{-\lambda}\frac{\lambda^i}{i!}$$

Poisson random variable satisfies property of probability since

$$\sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\left(\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\right) = e^{-\lambda}e^{\lambda} = 1$$

A Poisson random variable can be used to approximate a binomial random variable X with parameters (n, p) when n is large and p is small such that np is of moderate size that can be used as $\lambda = np$:

$$\begin{aligned} P\{X = k\} &= \binom{n}{k}p^k(1-p)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{1(1-1/n)\cdots(1-(k-1)/n)}{k!}\lambda^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &\approx \frac{\lambda^k}{k!}e^{-\lambda} \end{aligned}$$

Proposition 2.18. Let X be a Poisson random variable with parameter λ, then $E[X] = \lambda$ and $\mathrm{Var}(X) = \lambda$.

Proof. Note that

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-1)!} \\ &= \lambda\sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda\sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} = \lambda \end{aligned}$$

and by using similar approach as above,

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{(k-1)!} \\ &= \sum_{k=1}^{\infty} (k-1) e^{-\lambda} \frac{\lambda^k}{(k-1)!} + \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^k}{(k-1)!} \\ &= \lambda^2 + \lambda \end{aligned}$$

Therefore $\mathrm{Var}(X) = E[X^2] - E[X]^2 = \lambda$.

2.3 Properties of Expected Values

2.3.1 Expectation of Sums of Discrete Random Variables

Proposition 2.19. Let S be a finite or countably infinite sample space, and $p(s) = P(\{S\})$ for any $s \in S$, then for any random variable X on S,

$$E[X] = \sum_{s \in S} X(s) P(s)$$

Proof. Suppose the distinct values of X are x_i and $S_i = \{s \in S \mid X(s) = x_i\}$, then

$$\begin{aligned} E[X] &= \sum_i x_i P(\{X = x_i\}) \\ &= \sum_i x_i P(S_i) \\ &= \sum_i x_i \sum_{s \in S_i} p(s) \\ &= \sum_i \sum_{s \in S_i} X(s) p(s) \\ &= \sum_{s \in S} X(s) P(s) \end{aligned}$$

Proposition 2.20. Let $X_1, X_2, \cdots, X_n$ be discrete random variables on a finite or countably finite sample space S, then

$$E[X_1 + X_2 + \cdots + X_n] = \sum_{k=1}^{n} E[X_k]$$

Proof. By *Proposition 2.19*,

$$\begin{aligned} E[X_1 + X_2 + \cdots + X_n] &= \sum_{s\in S}(X_1(s) + X_2(s) + \cdots + X_n(s))p(s) \\ &= \sum_{k=1}^{n}\left(\sum_{s\in S} x_k(s)p(s)\right) \\ &= \sum_{k=1}^{n} E[X_k] \end{aligned}$$

2.3.2 Continuity of Probability

Definition 2.21. Let M be a collection of sets and $E_n \in M$ for all n, then M is said to be **closed under countable increasing unions**, denoted by $E_n \nearrow E$, if

$$E_{n+1} \supset E_n,\ E = \bigcup_{n=1}^{\infty} E_n$$

Similarly, M is said to be **closed under contable decreasing intersections**, denoted by $E_n \searrow E$, if

$$E_{n+1} \subset E_n,\ E = \bigcap_{n=1}^{\infty} E_n$$

With the defintion above, below is the continuity property of probability:

Proposition 2.22. Let M be a collection of sets and $E_n \in M$ for all n, then if $E_n \nearrow E$ or $E_n \searrow E$, $\lim P(E_n) = P(E)$.

2.3.3 Cumulative Distribution Function

Definition 2.23. Let X be a random variable on a sample space S, then the **cumulative distribution function** of X, denoted by F, is a function that maps from $\mathbb{R}$ to $\mathbb{R}$ such that

$$F(b) = P(\{X \leq b\})$$

Proposition 2.24. A cumulative distribution function F has the following properties:

(a) F is a nondecreasing function. In other words, for $a < b$, $F(a) \leq F(b)$.

(b) When b tends to $+\infty$ and $-\infty$ respectively,

$$\lim_{b\to+\infty} F(b) = 1, \quad \lim_{b\to-\infty} F(b) = 0$$

(c) F is right continuous. In other words,

$$\lim_{b_n\to b+} F(b_n) = F(b)$$

Proof. *Part (a)* If $a < b$, then $\{X \le a\} \subset \{X \le b\}$, so $F(a) \le F(b)$.

Part (b) If $b_n \nearrow \infty$, then

$$\{X \le b_n\} \nearrow \{X < \infty\} = S$$

so $F(b_n) \to 1$ by *Proposition 2.22*. Similarly, if $b_n \searrow -\infty$, then

$$\{X \le b_n\} \searrow \{X = -\infty\} = \phi$$

so $F(b_n) \to 0$ by *Proposition 2.22*.

Part (c) If $b_n \searrow b$, then

$$\{X \le b_n\} \searrow \{X \le b\}$$

so $F(b_n) \to F(b)$ by *Proposition 2.22* and shows that F is right continuous.

Note that F is not left continuous. In the discrete case,

$$P(\{X = b\}) = F(b) - \lim_{b_n\nearrow b} F(b_n) = F(b) - F(b-)$$

That is, if $b_n \nearrow b$, then

$$\{X \le b_n\} \nearrow \{X \le b\}$$

so

$$P(\{X = b\}) = P(\{X \le b\}) - P(\{X < b\})$$

2.4 Continuous Random Variables

2.4.1 Definition of Continuous Random Variables

Definition 2.25. Let f be a nonnegative function defined on $(-\infty, \infty)$, then X is called a **continuous random variable** if

$$P(\{X \in B\}) = \int_B f(x)\,\mathrm{d}x$$

for all measurable sets $B \subset (-\infty, \infty)$.

By measurable sets, it represents all intervals and countable unions or intersections of intervals. Note that

$$P(\{a \leq X \leq b\}) = \int_a^b f(x)\,\mathrm{d}x$$

which is the area of shaded region under $f(x)$.

Definition 2.26. Let f be a nonnegative function defined on $(-\infty, \infty)$ and X is a continuous random variable, then f is called a **probability density function** of X if

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

Example 2.27. Let X be a continuous random variable with probability density function

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{if } x \in (0, 2) \\ 0 & \text{otherwise} \end{cases}$$

Find the value of C, and $P(\{X \geq 1\})$.

Answer. By *Definition 2.26*,

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x &= 1 \\ \int_0^2 C(4x - 2x^2) &= 1 \\ C\left[2x^2 - \frac{2}{3}x^3\right]_0^2 &= 1 \\ \frac{8}{3}C &= 1 \\ C &= \frac{3}{8} \end{aligned}$$

Therefore

$$\begin{aligned} P(\{X \geq 1\}) &= \int_1^\infty f(x)\,\mathrm{d}x \\ &= \int_1^2 \frac{3}{8}(4x - 2x^2)\,\mathrm{d}x \\ &= \frac{3}{8}\left[2x^2 - \frac{2}{3}x^3\right]_1^2 = \frac{1}{2} \end{aligned}$$

Finally note that in the continuous case,

$$P(\{X = a\}) = \int_a^a f(x)\,\mathrm{d}x = 0$$

for any $a \in \mathbb{R}$, so $P(\{a \leq X \leq b\}) = P(\{a < X < b\})$.

2.4.2 Expectation of Continuous Random Variables

Definition 2.28. Let X be a continuous random variable with probability density function f, then the expectation of X is defined as

$$E[X] = \int_{-\infty}^\infty x f(x)\,\mathrm{d}x$$

Recall that in the discrete case, $E[X] = \sum xP(\{X = x\})$. In order to apply this equation for continuous random variables, set a partition of $(-\infty, \infty)$ by $(x_n)_{n=-\infty}^\infty$ such that $x_{n+1} - x_n = \Delta x$, then

$$\begin{aligned} &\sum_n x_n P(\{x_n < X < x_{n+1}\}) \\ =&\sum_n x_n \int_{x_n}^{x_n + \Delta x} f(x)\,\mathrm{d}x \\ \approx&\sum_n x_n (f(x_n)\Delta x) \end{aligned}$$

When $\Delta x \to 0$,

$$\lim_{\Delta x \to 0} \sum_n x_n (f(x_n)\Delta x) = \int_{-\infty}^\infty x f(x)\,\mathrm{d}x$$

Example 2.29. A continuous random variable X is said to be **uniformly distributed** on $[0, 1]$ if it has the density

$$f(x) = \begin{cases} 1 & \text{if } x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

Find $E[X]$.

Answer. Note that

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x = \int_0^1 x\,\mathrm{d}x = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

Proposition 2.30. Let Y be a nonnegative continuous random variable, then

$$E[Y] = \int_0^{\infty} P(\{Y > y\})\,\mathrm{d}y$$

Proof. Let f be the density of Y, then

$$\begin{aligned}\int_0^{\infty} P(\{Y > y\})\,\mathrm{d}y &= \int_0^{\infty} \left(\int_y^{\infty} f(x)\,\mathrm{d}x\right)\,\mathrm{d}y \\ &= \int_0^{\infty} \left(\int_0^{\infty} \mathbf{1}_{\{x>y\}} f(x)\,\mathrm{d}x\right)\,\mathrm{d}y\end{aligned}$$

where

$$\mathbf{1}_{\{x>y\}} = \begin{cases} 1 & \text{if } x > y \\ 0 & \text{otherwise} \end{cases}$$

By Fubini's theorem,

$$\begin{aligned}\int_0^{\infty} \left(\int_0^{\infty} \mathbf{1}_{\{x>y\}} f(x)\,\mathrm{d}x\right)\,\mathrm{d}y &= \int_0^{\infty} \left(\int_0^{\infty} \mathbf{1}_{\{x>y\}} f(x)\,\mathrm{d}y\right)\,\mathrm{d}x \\ &= \int_0^{\infty} f(x) \left(\int_0^{\infty} \mathbf{1}_{\{x>y\}}\,\mathrm{d}y\right)\,\mathrm{d}x \\ &= \int_0^{\infty} f(x)(x)\,\mathrm{d}x = E[Y]\end{aligned}$$

The following propostion is the general case about the expectation of continuous random variable:

Proposition 2.31. Let X be a continuous random variable and g be a real-valued function, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x$$

Proof. By *Proposition 2.30*,

$$\begin{aligned} E[g(x)] &= \int_0^\infty P(\{g(X) > y\})\,\mathrm{d}y \\ &= \int_0^\infty \left(\int_{-\infty}^\infty \mathbf{1}_{\{g(x)>y\}} f(x)\,\mathrm{d}x \right) \mathrm{d}y \\ &= \int_{-\infty}^\infty \left(\int_0^\infty \mathbf{1}_{\{g(x)>y\}} f(x)\,\mathrm{d}y \right) \mathrm{d}x \\ &= \int_{-\infty}^\infty f(x) \left(\int_0^\infty \mathbf{1}_{\{g(x)>y\}}\,\mathrm{d}y \right) \mathrm{d}x \\ &= \int_{-\infty}^\infty f(x)g(x)\,\mathrm{d}x \end{aligned}$$

Proposition 2.32. Let X be a continuous random variable with density f, then

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

where $\mu = E[X]$.

Proof. Note that

$$\begin{aligned} \mathrm{Var}(X) &= \int_{-\infty}^\infty (x^2 + 2x\mu + \mu^2) f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^\infty x^2 f(x)\,\mathrm{d}x + \mu \int_{-\infty}^\infty 2x f(x)\,\mathrm{d}x + \mu^2 \int_{-\infty}^\infty f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^\infty x^2 f(x)\,\mathrm{d}x - 2\mu^2 + \mu^2 \\ &= \int_{-\infty}^\infty x^2 f(x)\,\mathrm{d}x - \mu^2 \end{aligned}$$

2.5 Common Types of Continuous Random Variables

2.5.1 Uniform Random Variables

Definition 2.33. Let X be a continuous random variable, then X is a **uniform random variable** on $[a,b]$ if it has density

$$f(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Example 2.34. Let X be a uniform random variable. Calculate $E[X]$ and $\mathrm{Var}(X)$.

Answer. Note that

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \\ &= \int_a^b x\left(\frac{1}{b-a}\right)\,\mathrm{d}x \\ &= \frac{1}{b-a}\left[\frac{1}{2}x^2\right]_a^b \\ &= \frac{a+b}{2} \end{aligned}$$

and

$$\begin{aligned} &\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x \\ &= \int_a^b x^2\left(\frac{1}{b-a}\right)\,\mathrm{d}x \\ &= \frac{1}{b-a}\left[\frac{1}{3}x^3\right]_a^b \\ &= \frac{a^2+ab+b^2}{3} \end{aligned}$$

Therefore

$$\mathrm{Var}(X) = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(a-b)^2}{12}$$

For continuous random variables, it is also important to know the following definition:

Definition 2.35. Let X be a continuous random variable with density f, then **cumulative distribution function** of X, denoted by F_X, is defined as

$$F_X(b) = \int_{-\infty}^{b} f(x)\,\mathrm{d}x$$

Proposition 2.36. Let X be a continuous random variable with density f and cumulative distribution function F_X. If f is continuous at b, then $F_X'(b) = f(b)$.

Proof. For $u \in \mathbb{R}$ such that $u \neq 0$,

$$\begin{aligned} \frac{F_X(b+u) - F_X(b)}{u} &= \frac{1}{u}\left(\int_{-\infty}^{b+u} f(x)\,\mathrm{d}x - \int_{-\infty}^{b} f(x)\,\mathrm{d}x\right) \\ &= \frac{1}{u}\int_b^{b+u} f(x)\,\mathrm{d}x \end{aligned}$$

Since f is continuous at b, so $f(x) - f(b)$ is close at 0 when x is close to b. Hence when $u \to 0$,

$$\frac{1}{u}\int_b^{b+u} f(x)\,\mathrm{d}x \to f(b)$$

2.5.2 Normal Random Variables

Definition 2.37. Let $\mu \in \mathbb{R}$ and $\sigma > 0$, then a **normal random variable** with μ and σ^2 is a continuous random variable with density function

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

for $x \in \mathbb{R}$.

Note that it is not explicit to show that f is a density function. By substituting $y = (x-\mu)/\sigma$,

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,\mathrm{d}y \\ &= \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,\mathrm{d}y\end{aligned}$$

Let

$$I = \int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y$$

such that the problem reduces to a special case in double integration. Since

$$\begin{aligned}I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^{\infty}\int_0^{2\pi} e^{-r^2/2} r\,\mathrm{d}\theta\,\mathrm{d}r \\ &= 2\pi\int_0^{\infty} r e^{-r^2/2}\,\mathrm{d}r \\ &= 2\pi\left[-e^{-r^2/2}\right]_0^{\infty} = 2\pi\end{aligned}$$

Then $I = \sqrt{2\pi}$ and

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

Definition 2.38. A normal random variable X is said to be **standard** if it has parameters $\mu = 0$ and $\sigma^2 = 1$.

Proposition 2.39. Let X be a normal random variable with parameters μ and σ^2, and $Z = (X - \mu)/\sigma$, then Z is the standard normal random variable.

Proof. The cumulative distribution of Z is

$$\begin{aligned} F_Z(b) &= P(\{Z \leq b\}) = P\left(\left\{\frac{X-\mu}{\sigma} \leq b\right\}\right) \\ &= P(\{X \leq \sigma b + \mu\}) \\ &= \int_{-\infty}^{\sigma b+\mu} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, \mathrm{d}y \end{aligned}$$

Differentiate both sides gives

$$F_Z'(b) = \frac{1}{2\pi} e^{-b^2/2} = f_Z(b)$$

Hence the density of Z is the density of standard normal random variable.

The next theorem will show that parameters of normal random variables represent expectations and variances.

Proposition 2.40. Let Z be a standard normal random variable, then $E[Z] = 0$ and $\mathrm{Var}(X) = 1$.

Proof. Note that

$$E[Z] = \int_{-\infty}^{\infty} x \left(\frac{1}{\sqrt{2\pi}}\right) e^{-x^2/2} \, \mathrm{d}x = \left[-\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right]_{-\infty}^{\infty} = 0$$

and

$$\begin{aligned} E[Z^2] &= \int_{-\infty}^{\infty} x^2 \left(\frac{1}{\sqrt{2\pi}}\right) e^{-x^2/2} \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}} \left(-e^{-x^2/2}\right)' \mathrm{d}x \\ &= \left[\frac{x}{\sqrt{2\pi}}(-e^{-x^2/2})\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}(-e^{-x^2/2}) \, \mathrm{d}x \\ &= 0 - (-1) = 1 \end{aligned}$$

Hence $\mathrm{Var}(X) = E[Z^2] - E[Z]^2 = 1$.

Proposition 2.41. Let $Y = aX + b$ where $a, b \in \mathbb{R}$ and X is a continuous random variable, then $E[Y] = aE[X] + b$ and $\mathrm{Var}(Y) = a^2\mathrm{Var}(X)$.

Proof. Let f be the density function of X, then

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} (ax+b) f(x)\,\mathrm{d}x \\ &= a\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x + b\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \\ &= aE[X] + b \end{aligned}$$

and

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} (ax+b)^2 f(x)\,\mathrm{d}x \\ &= a^2\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x + 2ab\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x + b^2\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \\ &= a^2E[X^2] + 2abE[X] + b^2 \end{aligned}$$

Therefore

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

Theorem 2.42. Let X be a normal random variable with parameters μ and σ^2, then $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Proof. Let $Z = (X-\mu)/\sigma$ be the standard normal random variable. By *Proposition 2.40* and *Proposition 2.41*, $E[X] = \sigma E[Z] + \mu = \mu$ and $\mathrm{Var}(X) = \sigma^2\mathrm{Var}(Z) = \sigma^2$.

An important property of normal distribution is that it can be used to approximate the binomial random variable with parameters (n, p) where n is large and p is fixed. This is similar to approximation using Poisson random variable, but both approximation have different requirements. Below is the **de Moivre-Laplace theorem** for such approximation:

Theorem 2.43. Let $0 < p < 1$ be a fixed value and X_n be the binomial random variable with parameters (n, p), then for $a, b \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\left\{a \le \frac{X_n - np}{\sqrt{np(1-p)}} \le b\right\}\right) = P(\{a \le Z \le b\}) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x$$

In other words, de Moivre-Laplace theorem shows that $(X_n - np)/\sqrt{np(1-p)}$ has an approximate normal distribution. The theorem is also known as a special case of the central limit theorem.

Definition 2.44. Let Z be the standard normal random variable and $a \in \mathbb{R}$, then Φ is a **cumulative distribution function** such that

$$\Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x = P(\{Z \leq a\})$$

Geometrically, Φ is the area under the standard normal curve to the left of a random variable X.

Example 2.45. Let X be a binomial random variable with parameters $(40, 1/2)$. Find $P(\{X = 20\})$.

Answer. Note that the exact value of the probability is

$$P(\{X = 20\}) = \binom{40}{20}\left(\frac{1}{2}\right)^{40} \approx 0.1254$$

In order to apply *Theorem 2.44*, first a continuity correction is applied, which is

$$P(\{X = 20\}) = P(\{19.5 \leq X \leq 20.5\})$$

then

$$\begin{aligned} P(\{19.5 \leq X \leq 20.5\}) &= P\left(\left\{\frac{19.5-20}{\sqrt{10}} \leq \frac{X-20}{\sqrt{10}} \leq \frac{20.5-20}{\sqrt{10}}\right\}\right) \\ &\approx P(\{-0.16 \leq Z \leq 0.16\}) \\ &= 2\left(\Phi(0.16) - \frac{1}{2}\right) \end{aligned}$$

Finally, by checking values of Φ in a table, $\Phi(0.16) = 0.5636$ and $P(\{X = 20\}) = 0.1272$.

Example 2.46. Let X be a normal random variable with parameters $(10, 36)$. Find $P(\{7 \leq X \leq 16\})$ given that $\Phi(1) = 0.8413$ and $\Phi(0.5) = 0.6915$.

Answer. Let $Z = (X-\mu)/\sigma = (X-10)/6$ such that Z is a standard normal random variable, then

$$\begin{aligned} P(\{7 \leq X \leq 16\}) &= P(\{-0.5 \leq Z \leq 1\}) \\ &= \Phi(1) - \Phi(-0.5) \\ &= \Phi(1) - (1 - \Phi(0.5)) = 0.3085 \end{aligned}$$

2.5.3 Exponential Random Variables

Definition 2.47. Let $\lambda > 0$. An **exponential random variable** X with parameter λ is a random variable with density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Example 2.48. Let X be an exponential random variable with parameter λ. Calculate $E[X]$, $E[X^n]$ and $\mathrm{Var}(X)$.

Answer. Let $n \geq 1$, then

$$\begin{aligned} E[X^n] &= \int_{-\infty}^{\infty} x^n f(x)\,\mathrm{d}x = \int_0^{\infty} x^n \lambda e^{-\lambda x}\,\mathrm{d}x \\ &= \int_0^{\infty} x^n (-e^{-\lambda x})'\,\mathrm{d}x \\ &= [-x^n e^{-\lambda x}]_0^{\infty} - \int_0^{\infty} n x^{n-1}(-e^{-\lambda x})\,\mathrm{d}x \\ &= \frac{n}{\lambda}\int_0^{\infty} x^{n-1}\lambda e^{-\lambda x}\,\mathrm{d}x = \frac{n}{\lambda}E[X^{n-1}] \end{aligned}$$

Hence

$$E[X^n] = \frac{n!}{\lambda^n}E[X^0] = \frac{n!}{\lambda^n}$$

which in particular $E[X] = 1/\lambda$ and $E[X^2] = 2/\lambda^2$. Therefore $\mathrm{Var}(X) = E[X^2] - E[X]^2 = 1/\lambda^2$.

Example 2.49. Suppose the length of a phone call (in minutes) is an exponential random variable with parameter 1/10. If Amy arrives immediately right ahead of Bob at a public telephone booth, find the probability that Bob has to wait for between 10 and 20 minutes.

Answer. Note that

$$P(\{10 \leq X \leq 20\}) = \int_{10}^{20} \lambda e^{-\lambda x}\,\mathrm{d}x = e^{-1} - e^{-2}$$

2.6 Distribution of Function of Continuous Random Variables

2.6.1 Density of Function of Continuous Random Variables

Let X be a continuous random variable with density f_X, and $g : \mathbb{R} \to \mathbb{R}$ be a real-valued function. Further let $Y = g(X)$ is then a **function of continuous random variable**. The discussion is to figure out a general formula for density of Y, but before that, consider the following special cases:

Example 2.50. Let X be a continuous random variable with density f_X and $Y = X^2$ be a function of X. Find the density of Y.

Answer. By comparing the cumulative distribution function of Y,

$$\begin{aligned} F_Y(y) &= P(\{Y \le y\}) = P(\{X^2 \le y\}) \\ &= \begin{cases} P(\{-\sqrt{y} \le X \le \sqrt{y}\}) & \text{if } y \ge 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Hence if $y \ge 0$,

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

Taking derivative with respect to y gives

$$f_Y(y) = \begin{cases} f_X(\sqrt{y})(1/2\sqrt{y}) - f_X(-\sqrt{y})(-1/2\sqrt{y}) & \text{if } y > 0 \\ 0 & \text{if } y < 0 \end{cases}$$

Example 2.51. Let X be an exponential random variable with parameter λ, and $Y = 1/X$ be a function of X. Find the probability density function of Y.

Answer. Note that $P(\{X \le 0\}) = 0$ implies $P(\{Y \le 0\}) = 0$. After that, the cumulative distribution function of Y is

$$F_Y(y) = P(\{Y \le y\}) = \begin{cases} P(\{1/X \le y\}) & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Hence when $y > 0$,

$$F_Y(y) = P(\{X \ge 1/y\}) = 1 - F_X\left(\frac{1}{y}\right)$$

Taking derivative with respect to y gives

$$f_Y(y) = \begin{cases} f_x(1/y)(1/y^2) = \lambda \exp(-\lambda/y)(1/y^2) & \text{if } y > 0 \\ 0 & \text{if } y < 0 \end{cases}$$

Below is a general formula for density of function of continuous random variables:

Theorem 2.52. Let X be a continuous random variable with density f_X, $g : \mathbb{R} \to \mathbb{R}$ be a strictly monotone and differentiable function, and $Y = g(X)$ is a function of X, then

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\,|(g^{-1}(y))'| & \text{if } y = g(x) \text{ for some } x \\ 0 & \text{otherwise} \end{cases}$$

Proof. Assume g is strictly increasing, then

$$F_Y(y) = P(\{Y \leq y\}) = P(\{g(X) \leq y\}) = F_X(g^{-1}(y))$$

and differentiate with respect to y gives

$$f_Y(y) = f_X(g^-1(y))(g^{-1}(y))'$$

and absolute value is taken because $f_Y(y) \geq 0$.

3 Joint Distributions

3.1 Introduction to Joint Distributions

3.1.1 Joint Cumulative Distributions

Definition 3.1. Let X and Y be two random variables on a sample space, then a **joint cumulative distribution** of X and Y, denoted by F, is defined by

$$F(a,b) = P(\{X \leq a, Y \leq b\})$$

Let F_X and F_Y be the cumulative distribution function of X and Y respectively, then F_X and F_Y are determined by F. This is because

$$\begin{aligned} F_X(a) &= P(\{X \leq a\}) = P(\{X \leq a, Y \leq \infty\}) \\ &= P\left(\lim_{b\to\infty} \{X \leq a, Y \leq b\}\right) \\ &= \lim_{b\to\infty} P((X \leq a, Y \leq b)) = \lim_{b\to\infty} F(a,b) \end{aligned}$$

In other words, $F_X(a)$ can be represented as $F(a,\infty)$, and similarly, $F_Y(b)$ can be represented as $F(\infty,b)$. Usually, F_X and F_Y are called the **marginal distributions** of X and Y respectively.

Also, theoretically all statements about X and Y are determined by the joint distribution of X and Y.

Example 3.2. Let F be the joint cumulative distribution function of X and Y. Find $P(\{X > a, Y > b\})$.

Answer. Note that

$$\begin{aligned} P(\{X > a, Y > b\}) &= P(\{X > a\} \cap \{Y > b\}) \\ &= 1 - P((\{X > a\} \cap \{Y > b\})^c) \\ &= 1 - P(\{X \leq a\} \cup \{Y \leq b\}) \\ &= 1 - P(\{X \leq a\}) - P(\{Y \leq b\}) + P(\{X \leq a\} \cap \{Y \leq b\}) \\ &= 1 - F(a,\infty) - F(\infty,b) + F(a,b) \end{aligned}$$

3.1.2 Discrete Joint Distributions

When X and Y are both discrete, the joint probability mass function p can be defined by

$$p(x,y) = P\{X = x, Y = y\}$$

where (x, y) are possible values of (X, Y), and therefore

$$F(a,b) = \sum_{\substack{(x,y)\,|\,p(x,y)>0 \\ x\le a, y\le b}} p(x,y)$$

Consequently,

$$p_X(a) = \sum_{y\,|\,p(a,y)>0} p(a,y),\ p_Y(b) = \sum_{x\,|\,p(x,b)>0} p(x,b)$$

3.1.3 Continuous Joint Distributions

Definition 3.3. Let X and Y be continuous random variables, then X and Y are said to be **jointly continuous** if there exists a function $f : \mathbb{R}^2 \to [0, \infty)$ such that

$$P(\{(X,Y) \in C\}) = \iint_C f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

where C is a measurable set in $\mathbb{R}^2$.

For instance, when C is the countable intersection or union of ractangles $[a, b] \times [c, d]$, C is a measurable set in $\mathbb{R}^2$.

Example 3.4. Suppose X and Y have a joint density function

$$f(x,y) = \begin{cases} 12xy(1-x) & \text{if } 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $P(\{X \le 1/2, Y \le 1/2\})$ and $P(\{X \le 1/2\})$.

Answer. Note that

$$\begin{aligned} P(\{X \le 1/2, Y \le 1/2\}) &= P(\{(X,Y) \in (-\infty, 1/2) \times (-\infty, 1/2)\}) \\ &= \int_{-\infty}^{1/2} \int_{-\infty}^{1/2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^{1/2} \int_0^{1/2} 12xy(1-x)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^{1/2} \left[12y\left(\frac{x^2}{2} - \frac{x^3}{3}\right)\right]_0^{1/2} \mathrm{d}y \\ &= \int_0^{1/2} y\,\mathrm{d}y \\ &= \left[\frac{y^2}{2}\right]_0^{1/2} = \frac{1}{8} \end{aligned}$$

and similarly,

$$P(\{X \leq 1/2\}) = \int_0^1 \int_0^{1/2} 12xy(1-x)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}$$

Example 3.5. Suppose X and Y have a joint density

$$f(x,y) = \begin{cases} e^{-(x+y)} & \text{if } 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Find the probability density function of X/Y.

Answer. Notice that f is nonzero only on $(0,\infty) \times (0,\infty)$, or in other words, when x and y are both positive. Assume X and Y always take positive values, then X/Y is also positive. For any $a > 0$,

$$\begin{aligned} F_{X/Y}(a) &= P\left(\left\{\frac{X}{Y} \leq a\right\}\right) = P(\{X \leq aY\}) \\ &= \iint\limits_{(x,y)\,|\,x \leq ay} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint\limits_{\substack{(x,y)\,|\,x \leq ay \\ x>0, y>0}} e^{-(x+y)}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^\infty \int_0^{ay} e^{-(x+y)}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^\infty e^{-y}\left[-e^{-x}\right]_0^{ay}\,\mathrm{d}y \\ &= \int_0^\infty e^{-y}(1-e^{-ay})\,\mathrm{d}y \\ &= \left[-e^{-y} + \frac{e^{-(1+a)y}}{1+a}\right]_0^\infty \\ &= 1 - \frac{1}{1+a} \end{aligned}$$

Hence

$$f_{X/Y}(a) = \begin{cases} 1/(1+a)^2 & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}$$

If X and Y are jointly continuous with a joint density f, then the cumulative distribution function

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

for any $a, b \in \mathbb{R}$. Similarly, if f is continuous at (a, b) then

$$\frac{\partial^2 F(a,b)}{\partial a \partial b} = f(a,b)$$

This is because if

$$g(y) = \int_{-\infty}^{a} f(x,y)\,\mathrm{d}x$$

then

$$F(a,b) = \int_{-\infty}^{b} g(y)\,\mathrm{d}y$$

and by Fundamental Theorem of Calculus,

$$\frac{\partial F(a,b)}{\partial b} = g(b) = \int_{-\infty}^{a} f(x,b)\,\mathrm{d}x$$

and

$$\frac{\partial^2 F(a,b)}{\partial a \partial b} = f(a,b)$$

In the above joint continuous case,

$$f_X(a) = \int_{-\infty}^{\infty} f(a,y)\,\mathrm{d}y,\ f_Y(b) = \int_{-\infty}^{\infty} f(x,b)\,\mathrm{d}x$$

because

$$F_X(a) = P(\{X \le a, Y \le \infty\}) = \int_{-\infty}^{a}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x$$

and using the same technique as above,

$$f_X(a) = \int_{-\infty}^{\infty} f(a,y)\,\mathrm{d}y$$

3.2 Independence of Random Variables

3.2.1 Independence of Two Random Variables

Recall that two events E and F are independent if $P(E \cap F) = P(E)P(F)$. Below is the definition about independence of random variables:

Definition 3.6. Let X and Y be random variables, then they are **independent** if

$$P(\{X \in A, Y \in B\}) = P(\{X \in A\})P(\{Y \in B\})$$

for all measurable sets $A, B \subset \mathbb{R}$.

In other words, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent for all A and B. Equivalently, X and Y are said to be independent if $F(a,b) = F_X(a)F_Y(b)$ for all $a, b \in \mathbb{R}$.

3.2.2 Independence of Discrete Random Variables

Proposition 3.7. Let X and Y be discrete random variables, then X and Y are independent if and only if $p(x,y) = p_X(x)p_Y(y)$.

Proof. ($\Rightarrow$) Suppose X and Y are independent, then

$$p(x,y) = P(\{X = x, Y = y\}) = P(\{X = x\})P(\{Y = y\}) = p_X(x)p_Y(y)$$

($\Leftarrow$) Suppose $p(x,y) = p_X(x)p_Y(y)$ holds for all x and y, then

$$\begin{aligned} F(a,b) &= P(\{X \leq a, Y \leq b\}) \\ &= \sum_{x \leq a, y \leq b} p(x,y) \\ &= \sum_{x \leq a, y \leq b} p_X(x)p_Y(y) \\ &= \left(\sum_{x \leq a} p_X(x)\right)\left(\sum_{y \leq b} p_Y(y)\right) \\ &= F_X(a)F_Y(b) \end{aligned}$$

which means X and Y are independent.

3.2.3 Independence of Continuous Random Variables

Proposition 3.8. If X and Y are jointly continuous with a density $f(x,y)$, then X and Y are independent if and only if $f(x,y) = f_X(x)f_Y(y)$.

Proof. ($\Rightarrow$) If X and Y are independent, then $F(a,b) = F_X(a)F_Y(b)$ for any $a, b \in \mathbb{R}$. Taking partial derivatives gives

$$f(a,b) = \frac{\partial^2 F(a,b)}{\partial a \partial b} = F_X'(a)F_Y'(b) = f_X(a)f_Y(b)$$

($\Leftarrow$) If $f(x,y) = f_X(x)f_Y(y)$ holds for all x and y, then

$$\begin{aligned} F(a,b) &= \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{b}\int_{-\infty}^{a} f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \left(\int_{-\infty}^{b} f_Y(y)\,\mathrm{d}y\right)\left(\int_{-\infty}^{a} f_X(x)\,\mathrm{d}x\right) \\ &= F_X(a)F_Y(b) \end{aligned}$$

which means X and Y are independent.

Example 3.9. Suppose X and Y have a joint density $f(x,y) = 24xy$ if $0 < x < 1$, $0 < y < 1$ and $0 < x + y < 1$. Determine whether X and Y are independent.

Answer. For $0 < a < 1$,

$$f_X(a) = \int_{-\infty}^{\infty} f(a,y)\,\mathrm{d}y = \int_{0}^{1-a} 24ay\,\mathrm{d}y = 12a(1-a)^2$$

which means $f_Y(b) = 12b(1-b)^2$. Therefore $f(a,b) \neq f_X(a)f_Y(b)$ and X and Y are not independent.

The following example is called Buffon's needle problem, and it states as follows:

Example 3.10. A table is ruled with equidistant parallel lines, a distance D apart. A needle of length L where $L \leq D$ is randomly thrown on the table. Find the probability that the needle with intersect one of the lines.

Answer. Let O be the center of the needle, X be the shortest distance between O and the nearest parallel line, and θ be the angle between the needle and the vertical direction (perpendicular to parallel line). Note that $0 \leq X \leq D/2$ and $0 \leq \theta \leq \pi/2$.

Suppose X and θ has a uniform distribution on $[0, D/2]$ and $[0, \pi/2]$ respectively, then the needle intersects one of the parallel lines if and only if $X \leq (L/2)\cos(\theta)$. Further assume that X and θ are independent, then

$$
\begin{aligned}
P\left(\left\{X \le \frac{1}{2}L\cos(\theta)\right\}\right) &= \iint\limits_{\substack{(x,\theta)\in[0,D/2]\times[0,\pi/2] \\ x\le(L/2)\cos(\theta)}} f(x,\theta)\,\mathrm{d}x\,\mathrm{d}\theta \\
&= \iint\limits_{A} f_X(x) f_\theta(\theta)\,\mathrm{d}x\,\mathrm{d}\theta \\
&= \int_0^{\pi/2}\int_0^{(L/2)\cos(\theta)} \frac{2}{D}\frac{2}{\pi}\,\mathrm{d}x\,\mathrm{d}\theta \\
&= \int_0^{\pi/2} \frac{2L}{D\pi}\cos(\theta)\,\mathrm{d}\theta \\
&= \left[\frac{2L}{D\pi}\sin(\theta)\right]_0^{\pi/2} = \frac{2L}{D\pi}
\end{aligned}
$$

3.2.4 Independence of Multiple Random Variables

Note that the concepts of joint distribution and independence can be extended to more that two random variables. That is, if $X_1, X_2, \cdots, X_n$ are random variables on a sample space, then the joint cumulative distribution function is

$$F(a_1, a_2, \cdots, a_n) = P(\{X_1 \le a_1, X_2 \le a_2, \cdots, X_n \le a_n\})$$

Theorem 3.11. Let $X_1, X_2, \cdots, X_n$ be random variables on a sample space, then they are independent to each other if

$$P(\{X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n\}) = \prod_{k=1}^{n} P(\{X_k \in A_k\})$$

for measurable sets $A_1, A_2, \cdots, A_n \in \mathbb{R}$. Equivalently,

$$F(a_1, a_2, \cdots, a_n) = \prod_{k=1}^{n} F_{X_k}(a_k)$$

for all $a_1, a_2, \cdots, a_n \in \mathbb{R}$.

3.3 Sums of Independent Random Variables

3.3.1 Simple Sums of Independent Continuous Random Variables

Consider the following example:

Example 3.12. Let X, Y be independent. Suppose both of them have a uniform distribution on $[0, 1]$, calculate the distribution of $X + Y$.

Answer. Note that $X + Y \in [0, 2]$. For $0 \le a \le 2$,

$$P(\{X+Y\le a\}) = \iint\limits_{(x,y)\in[0,1]^2, x+y\le a} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

where

$$f(x,y) = f_X(x)f_Y(y) = \begin{cases} 1 & \text{if } 0<x<1, 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

Hence

$$\begin{aligned} P(\{X+Y\le a\}) &= \iint\limits_{(x,y)\in[0,1]^2, x+y\le a} 1\,\mathrm{d}x\,\mathrm{d}y \\ &= \begin{cases} \int_0^a \int_0^{a-x} 1\,\mathrm{d}y\,\mathrm{d}x & \text{if } 0<a<1 \\ \int_0^1 \int_0^{\min(a-x,1)} 1\,\mathrm{d}y\,\mathrm{d}x & \text{if } 1<a<2 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} a^2/2 & \text{if } 0<a<1 \\ \min(a-1/2,1) & \text{if } 1<a<2 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Now consider a more general method to compute the distribution of $X+Y$. Let X and Y be independent continuous random variables with density f_X and f_y respectively, and so $f(x,y) = f_X(x)f_Y(y)$. For $a\in\mathbb{R}$,

$$\begin{aligned} F_{X+Y}(a) &= P(\{X+Y\le a\}) \\ &= \iint\limits_{(x,y)\,|\,x+y\le a} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint\limits_{(x,y)\,|\,x+y\le a} f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)\,\mathrm{d}y \end{aligned}$$

For simplicity, denote $*$ such that for f and g,

$$f*g(a) = \int_{-\infty}^{\infty} f(a-y)g(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} f(y)g(a-y)\,\mathrm{d}y$$

Note that the density of $X+Y$ is

$$\begin{aligned} f_{X+Y}(a) &= \frac{\mathrm{d}F_{X+Y}(a)}{\mathrm{d}a} = \frac{\mathrm{d}}{\mathrm{d}a}\int_{-\infty}^{\infty} F_X(a-y)f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}a}F_X(a-y)f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)\,\mathrm{d}y = f_X * f_Y(a) \end{aligned}$$

Example 3.13. Suppose X and Y are independent normal random variables with parameters $(0,1)$ and (μ,σ^2). Show that $X+Y$ has a normal distribution with parameters $(0, 1+\sigma^2)$.

Answer. Density of $X+Y$

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(a-y)^2}{2}\right)\right)\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right)\right)\,\mathrm{d}y \\ &= \frac{1}{2\pi\sigma}\int_{-\infty}^{\infty}\exp\left(-\frac{(a-y)^2}{2}-\frac{y^2}{2\sigma^2}\right) \end{aligned}$$

Note that

$$\frac{(a-y)^2}{2}+\frac{y^2}{2\sigma^2} = \frac{(ky-a\sigma^2/k)^2}{2\sigma^2}+\frac{a^2}{2k^2}$$

where $k=\sqrt{\sigma^2+1}$. Hence

$$f_{X+Y}(a) = \frac{1}{2\pi\sigma}\exp\left(-\frac{a^2}{2k^2}\right)\int_{-\infty}^{\infty}\exp\left(\frac{(ky-a\sigma^2/k)^2}{2\sigma^2}\right)\,\mathrm{d}y$$

Let $z=(ky-a\sigma^2/k)^2/\sigma$, then

$$f_{X+Y}(a) = \frac{1}{2\pi k}\exp\left(-\frac{a^2}{2k^2}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{z^2}{2}\right)\,\mathrm{d}z = \frac{1}{\sqrt{2\pi}k}\exp\left(-\frac{a^2}{2k^2}\right)$$

which implies $X+Y$ is normal with parameters $(0,k^2)$.

3.3.2 Simple Sums of Independent Discrete Random Variables

If X and Y are independent and discrete random variables,

$$\begin{aligned} P(\{X+Y=a\}) &= \sum_x P(\{X=x, Y=y\}) \\ &= \sum_x P(\{X=x\})P(\{Y=a-x\}) \\ &= \sum_x p_X(x)p_Y(a-x) \end{aligned}$$

Example 3.14. Suppose X and Y are independent Poisson random variables with parameters λ_1 and λ_2 respectively. Find the distribution of $X+Y$.

Answer. Since both X and Y take values in $\{0,1,2,\cdots\}$, $X+Y$ also takes values in $\{0,1,2,\cdots\}$. For any integer $n \geq 0$,

$$\begin{aligned} P(\{X+Y=n\}) &= \sum_{k=0}^{n} P(\{X=k\})P(\{Y=n-k\}) \\ &= \sum_{k=0}^{n} \left(e^{-\lambda_1}\frac{\lambda_1^k}{k!}\right)\left(e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!}\right) \\ &= \frac{e^{-\lambda_1-\lambda_2}}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^k\lambda_2^{n-k} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}(\lambda_1+\lambda_2)^n \end{aligned}$$

which implies $X+Y$ has a Poisson distribution with parameter $\lambda_1+\lambda_2$.

3.4 Conditional Distributions

3.4.1 Definition of Conditional Distributions

Definition 3.15. Let X and Y be discrete random variables, then the **conditional probability mass function** of X given $Y=y$ is

$$\begin{aligned} p_{X|Y}(x,y) &= P\{X=x \mid Y=y\} \\ &= \frac{P(X=x, Y=y)}{P(Y=y)} \\ &= \frac{p(x,y)}{p_Y(y)} \end{aligned}$$

if $p_Y(y) \neq 0$.

Example 3.16. Let X and Y be two independent Poisson random variables with parameters λ_1 and λ_2. Calculate the conditional distribution of X given $X+Y=n$ for some fixed nonnegative integer n.

Answer. Note that

$$\begin{aligned} P\{X = k \mid X + Y = n\} &= \frac{P\{X = k, X + Y = n\}}{P\{X + Y = n\}} \\ &= \frac{P\{X = k\}\, P\{Y = n - k\}}{P\{X + Y = n\}} \\ &= \frac{(e^{-\lambda_1}\lambda_1^k/k!)(e^{-\lambda_2}\lambda_2^{n-k}/(n-k)!)}{e^{-\lambda_1-\lambda_2}(\lambda_1+\lambda_2)^n/n!} \\ &= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \end{aligned}$$

That is, the conditional distribution above is binomial with parameters n and $\lambda_1/(\lambda_1+\lambda_2)$.

Definition 3.17. Let X and Y be discrete random variables, then for $A \subset \mathbb{R}$, the **conditional probability** of X taking values in A given $Y = y$ is

$$P\{X \in A \mid Y = y\} = \int_A f_{X|Y}(x, y)\, \mathrm{d}x$$

From the definition above, for any $a \in A$,

$$F_{X|Y}(a, y) = P\{X \leq a \mid Y = y\} = \int_{-\infty}^{a} f_{X|Y}(x, y)\, \mathrm{d}x$$

First, note that if X and Y are independent, $f_{X|Y}(x, y) = f_X(x)$. Also, by the meaning of $Y = y$, one can consider the equation as

$$\begin{aligned} P\{X \in A \mid Y = y\} &= \lim_{\epsilon \to 0} P\{X \in A \mid y - \epsilon < Y < y + \epsilon\} \\ &= \lim_{\epsilon \to 0} \frac{P\{X \in A, y - \epsilon < Y < y + \epsilon\}}{P\{y - \epsilon < Y < y + \epsilon\}} \end{aligned}$$

Example 3.18. Suppose the joint density of random variables X and Y is given by

$$f(x, y) = \begin{cases} e^{-x/y}e^{-y}/y & \text{if } x > 0, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find $P\{X > 1 \mid Y = y\}$.

Answer. Note that

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x \\ &= \int_0^{\infty} e^{-x/y}e^{-y}/y\,\mathrm{d}x \\ &= \left[-e^{-x/y}e^{-y}\right]_0^{\infty} = e^{-y} \end{aligned}$$

Then

$$\begin{aligned} P\{X>1 \mid Y=y\} &= \int_1^{\infty} \frac{f(x,y)}{f_Y(y)}\,\mathrm{d}x \\ &= \int_1^{\infty} e^{-x/y}/y\,\mathrm{d}x \\ &= \left[-e^{-x/y}\right]_1^{\infty} = e^{-1/y} \end{aligned}$$

if $y>0$. Otherwise, $P\{X>1 \mid Y=y\}=0$.

3.4.2 Joint Distributions of Functions of Random Variables

Recall the Jacobian (determinant) of f such that $(x_1,x_2)\mapsto(g_1(x_1,x_2),g_2(x_1,x_2))$ is

$$J(x_1,x_2)=\det\left(\begin{bmatrix}\partial g_1/\partial x_1 & \partial g_1/\partial x_2\\ \partial g_2/\partial x_1 & \partial g_2/\partial x_2\end{bmatrix}\right)=\frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2}-\frac{\partial g_2}{\partial x_1}\frac{\partial g_1}{\partial x_2}$$

Let X_1 and X_2 be jointly continuous random variables with density $f_{X_1,X_2}(x_1,x_2)$. Further let $g_1,g_2:\mathbb{R}^2\to\mathbb{R}$ such that $Y_1=g_1(X_1,X_2)$ and $Y_2=g_2(X_1,X_2)$. With the new random variables, the objective is the find the joint distribution of Y_1 and Y_2.

Theorem 3.19. Let X_1, X_2, Y_1 and Y_2 be random variables, and g_1 and g_2 be mapping functions defined as above with the following assumptions:

(a) x_1 and x_2 can be solved in terms of y_1 and y_2.

(b) g_1 and g_2 have continuous partial derivatives and the Jacobian $J(x_1,x_2)\neq 0$.

Then Y_1 and Y_2 have joint density

$$f_{Y_1,Y_2}(y_1,y_2)=\frac{f_{X_1,X_2}(x_1,x_2)}{|J(x_1,x_2)|}$$

Consider the following example:

Example 3.20. Let X_1 and X_2 be jointly continuous random variables with density $f(x_1,x_2)$. Further let $Y_1=X_1+X_2$ and $Y_2=X_1-X_2$, find the joint density of Y_1 and Y_2.

Answer. Let $y_1=g_1(x_1,x_2)=x_1+x_2$ and $y_2=g_2(x_1,x_2)=x_1-x_2$, then

$x_1 = (y_1 + y_2)/2$ and $x_2 = (y_1 - y_2)/2$. Note that

$$J(x_1, x_2) = \det\left(\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right) = -2$$

Therefore

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{f(x_1, x_2)}{|J(x_1, x_2)|} = \frac{1}{2} f\left(\frac{x_1 + x_2}{2}, \frac{x_1 - x_2}{2}\right)$$

4 Other Properties of Probability

4.1 Properties of Expectations

4.1.1 Expectations of Functions and Sums of Random Variables

Recall that

$$E[X] = \sum_x xp(x)$$

for discrete case and

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x$$

for continuous case. No matter which case it is, the expectation of X represents a weighted average of all possible values of X.

Proposition 4.1. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a real-valued function, and X and Y be discrete random variables with a joint probability mass function $p(x, y)$, then

$$E[g(X, Y)] = \sum_x \sum_y g(x, y)p(x, y)$$

Proposition 4.2. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a real-valued function, and X and Y be continuous random variables with density $f(x, y)$, then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Proof. Assume g is nonnegative, then apply the formula for expectation to give

$$\begin{aligned}
E[g(X, Y)] &= \int_0^{\infty} P\{g(X, Y) > t\}\,\mathrm{d}t \\
&= \int_0^{\infty} \left(\iint_{g(x,y)>t} f(x, y)\,\mathrm{d}x\,\mathrm{d}y \right) \mathrm{d}t \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\int_0^{g(x,y)} f(x, y)\,\mathrm{d}t \right) \mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)\,\mathrm{d}x\,\mathrm{d}y
\end{aligned}$$

Corollary. Let $X_1, X_2, \cdots, X_n$ be random variables, then

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Proof. The proof is done by induction. Assume X and Y are jointly continuous random variables with density $f(x, y)$, then by *Proposition 4.2*,

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)\,\mathrm{d}x\,\mathrm{d}y + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}xf(x,y)\,\mathrm{d}y\right)\mathrm{d}x + \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}yf(x,y)\,\mathrm{d}x\right)\mathrm{d}y \\ &= \int_{-\infty}^{\infty}xf_X(x)\,\mathrm{d}x + \int_{-\infty}^{\infty}yf_Y(y)\,\mathrm{d}y = E[X]+E[Y] \end{aligned}$$

Then for any n, substitute $X = X_1 + X_2 + \cdots + X_{n-1}$ and $Y = X_n$ which will give the resulting equation.

4.2 Covariances

4.2.1 Definition of Covariances

Definition 4.3. Let X and Y be random variables, then the **covariance** of X and Y, denoted by $\mathrm{Cov}(X, Y)$, is defined by

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

Note that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$. Also, similar to variance that $\mathrm{Var}(X) = E[X^2] - E[X]^2$, another formula for covariance is $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$. This will be explained briefly in the next section.

4.2.2 Properties of Covariances

Proposition 4.4. Let X and Y be independent random variables, and $g, h : \mathbb{R} \to \mathbb{R}$ be real-valued functions, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Proof. For continuous case,

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \left(\int_{-\infty}^{\infty} g(x)f_X(x)\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} h(y)f_Y(y)\,\mathrm{d}y\right) \\ &= E[g(X)]E[h(Y)] \end{aligned}$$

Corollary. If X and Y are independent random variables, then $\mathrm{Cov}(X,Y) = 0$.

Proof. By *Proposition 4.4*,

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])] = E[X - E[X]]E[Y - E[Y]] = 0$$

An important note for the corollary above is that the inverse does not hold. That is, $\mathrm{Cov}(X,Y) = 0$ does not imply X and Y are independent. Consider the following example:

Example 4.5. Let X and Y be random variables where

$$P\{X = -1\} = P\{X = 0\} = P\{X = 1\} = \frac{1}{3}$$

and

$$Y = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases}$$

Since $E[X] = 0$ and $E[XY] = 0$, by the second formula of covariance, $\mathrm{Cov}(X,Y) = 0$. However, when substitute $(x,y) = (0,0)$,

$$P\{X = 0\}\,P\{Y = 0\} = \frac{1}{3}\left(1 - \frac{1}{3}\right) \neq 0 = P\{X = 0, Y = 0\}$$

implies X and Y are not independent.

Here is a standard list of properties of covariance:

Proposition 4.6. Let X and Y be random variables, then the following equation holds:

(a)

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

(b)
$$\mathrm{Cov}(X, X) = \mathrm{Var}(X)$$

(c) For any $a \in \mathbb{R}$,
$$\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y)$$

(d) If $X = X_1 + X_2 + \cdots + X_n$ and $Y = Y_1 + Y_2 + \cdots + Y_m$, then
$$\mathrm{Cov}(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$$

Corollary. Let $X_1, X_2, \cdots, X_n$ be random variables, then
$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)$$

Moreover, if $X_1, X_2, \cdots, X_n$ are piecewise independent, then
$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

4.2.3 Independent and Identically Distributed Random Variables

Definition 4.7. Let $X_1, X_2, \cdots, X_n$ be random variables, then they are **identically distributed** if they shares the same expected value μ and variance σ^2.

Below is an example which discuss sample mean and sample variance:

Example 4.8. Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed (IID) random variables with expected value μ and variance σ^2. The sample mean and sample variance, denoted by $\overline{X}$ and S^2 respectively, is defined by
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \; S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$$

(a) Find the variance of sample mean $\mathrm{Var}(\overline{X})$.

(b) Find the expected value of sample variance $E[S^2]$.

(c) Show that $\text{Cov}(X_i - \overline{X}, \overline{X}) = 0$.

Answer. Part (a) Note that

$$\begin{aligned}\text{Var}(\overline{X}) &= \text{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n^2}\text{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{\sigma^2}{n}\end{aligned}$$

Part (b) Note that

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2 \\ &= \sum_{i=1}^{n}\left((X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2\right) \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\overline{X} - \mu)^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\end{aligned}$$

Then

$$E[(n-1)S^2] = \sum_{i=1}^{n}E[(X_i - \mu)^2] - nE[(\overline{X} - \mu)^2] = (n-1)\sigma^2$$

which implies

$$E[S^2] = \frac{1}{n-1}E[(n-1)S^2] = \sigma^2$$

Part (c) Note that

$$\begin{aligned}\text{Cov}(X_i - \overline{X}, \overline{X}) &= \text{Cov}(X_i, \overline{X}) - \text{Var}(\overline{X}) \\ &= \text{Cov}\left(X_i, \frac{X_1 + X_2 + \cdots + X_n}{n}\right) - \frac{\sigma^2}{n} \\ &= \frac{1}{n}\sum_{j=1}^{n}\text{Cov}(X_i, X_j) - \frac{\sigma^2}{n} \\ &= \frac{1}{n}\text{Var}(X_i) - \frac{\sigma^2}{n} = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0\end{aligned}$$

4.3 Conditional Expectations

4.3.1 Definition of Conditional Expectations

Definition 4.9. Let X and Y be discrete random variables, then the **conditional expectation** of X given $Y = y$ is

$$E[X \mid Y = y] = \sum_x xP\{X = x \mid Y = y\}$$

provided that $P\{Y = y\} > 0$.

Definition 4.10. Let X and Y be continuous random variables with density $f(x, y)$, then the conditional expectation of X given $Y = y$ is

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x$$

where

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

provided that $f_Y(y) > 0$.

Example 4.11. Let X and Y be jointly continuous random variable with density

$$f(x, y) = \begin{cases} e^{-x/y}e^{-y}/y & \text{if } x, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Calculate $E[X \mid Y = y]$ given that $y > 0$.

Answer. Note that

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x \\ &= \int_0^{\infty} \frac{e^{-x/y}e^{-y}}{y}\,\mathrm{d}x \\ &= \left[-e^{x/y}e^{-y}\right]_0^{\infty} = e^{-y} \end{aligned}$$

where $y > 0$. This gives

$$f_{X|Y}(x \mid y) = \frac{e^{-x/y}e^{-y}/y}{e^{-y}} = \frac{e^{-x/y}}{y}$$

Therefore

$$\begin{aligned} E[X \mid Y = y] &= \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x \\ &= \int_0^{\infty} \frac{xe^{-x/y}}{y}\,\mathrm{d}x \\ &= \left[-xe^{-x/y}\right]_0^{\infty} - \int_0^{\infty} -e^{-x/y}\,\mathrm{d}x \\ &= 0 + \left[-ye^{-x/y}\right]_0^{\infty} = y \end{aligned}$$

where $y > 0$.

4.3.2 Law of Total Expectation

Assume $E[X \mid Y]$ is a function of Y by $y \mapsto E[X \mid Y = y]$, then the following theorem called **law of total expectation** (or **Adam's law**) can be applied:

Theorem 4.12. Let X and Y be random variables, then $E[X] = E[E[X \mid Y]]$.

Proof. Note that in discrete case,

$$\begin{aligned} E[E[X \mid Y]] &= \sum_y E[X \mid Y = y] p_Y(y) \\ &= \sum_y \sum_x xP\{X = x \mid Y = y\}\, p_Y(y) \\ &= \sum_y \sum_x xP\{X = x, Y = y\} \\ &= \sum_x \sum_y xP\{X = x, Y = y\} \\ &= \sum_x xP\{X = x\} = E[X] \end{aligned}$$

Example 4.13. A miner is trapped in a mine containing 3 doors. The first door leads to a tunnel that takes him to safety after 3 hours of travel, and the other two doors that takes him back to the mine after 5 and 7 hours of travel respectively. Assume the miner is equally likely to choose any door at all times, what is the expected length of time until he reaches safety?

Answer. Let X be the number of hours until the miner reaches safety, and Y be the door he choose in the first time. By *Theorem 4.12*,

$$\begin{aligned} E[X] &= E[E[X \mid Y]] \\ &= \sum_{i=1}^{3} E[X \mid Y = i] P\{Y = i\} \\ &= 3\left(\frac{1}{3}\right) + (5 + E[X])\left(\frac{1}{3}\right) + (7 + E[X])\left(\frac{1}{3}\right) \end{aligned}$$

By solving the equation above, $E[X] = 3 + 5 + 7 = 15$ hours.

4.4 Moment Generating Functions

4.4.1 Defintion of Moment Generating Functions

Definition 4.14. Let X be a random variable and $t \in \mathbb{R}$, then the **moment generating function** of X, denoted by $M_X(t)$, is defined as $M_X(t) = E[e^{tX}]$.

If the moment generating function is clear to represent a random variable, the notation becomes $M(t)$ for convenience. Note that

$$e^{tX} = \sum_{n=0}^{\infty} \frac{t^n}{n!} X^n$$

implies

$$M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} E[X^n]$$

where $E[X^n]$ is called the **n-th moment** of X. Moment generating functions may not be useful at first glance, but if $M_X(t)$ exists and is finite for all $-t_0 < t < t_0$ for some $t_0 > 0$, then $E[X^n] = M_X(0)$ for any positive integer n.

4.4.2 Examples of Moment Generating Functions

Consider the following examples of finding moment generating functions for some common distributions:

Example 4.15. Let X be binomial random variable with parameters (n, p), then

$$\begin{aligned} M(t) = E[e^{tX}] &= \sum_{k=0}^{n} e^{tk} P\{X = k\} \\ &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} = (pe^t - p + 1)^n \end{aligned}$$

Example 4.16. Let X be Poisson random variable with parameters λ, then

$$\begin{aligned} M(t) = E[e^{tX}] &= \sum_{k=0}^{\infty} e^{tk} P\{X = k\} \\ &= \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \sum_{k=0}^{\infty} e^{-\lambda} \frac{(\lambda e^t)^k}{k!} = \exp(\lambda(e^t - 1)) \end{aligned}$$

Example 4.17. Let Z be standard normal random variable, then

$$\begin{aligned} M(t) = E[e^{tZ}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz} e^{-z^2/2} \, \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t^2/2} e^{-(z-t)^2/2} \, \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t^2/2} e^{-(z-t)^2/2} \, \mathrm{d}(z-t) = e^{t^2/2} \end{aligned}$$

Example 4.18. Let X be normal random variable with mean μ and variance σ^2, then by *Example 4.17*,

$$\begin{aligned} M_X(t) = E[e^{tX}] &= E[e^{t(\mu+\sigma Z)}] = E[e^{t\mu} e^{t\sigma Z}] \\ &= e^{t\mu} E[e^{(t\sigma)Z}] = e^{t\mu} M_Z(t\sigma) \\ &= e^{t\mu} e^{(t\sigma)^2/2} = \exp\left(\frac{(\sigma t)^2}{2} + \mu t\right) \end{aligned}$$

4.4.3 Moment Generating Functions and Distributions

Theorem 4.19. Let X and Y be random variables, then if for any $t_0 > 0$ such that $M_X(t) = M_Y(t) = c$ for $t \in (-t_0, t_0)$ where c is a finite number, then X and Y have the same distribution.

In other words, the moment generating function of any random variable determines its distribution.

Proposition 4.20. Let X and Y be independent random variables, then $M_{X+Y}(t) = M_X(t) + M_Y(t)$.

Proof. Note that

$$M_{X+Y}(t) = E[e^{tX+tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$$

4.5 Limiting Theorems

4.5.1 Markov's Inequality and Chebyshev's Inequality

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and identically distributed random variables, and the objective is to figure out the limiting behaviour of $(X_1+X_2+\cdots+X_n)/n$ when n tends to infinity. Below are **Markov's inequality** and **Chebyshev's inequality** in order to solve the problem above:

Theorem 4.21. Let X be a nonnegative random variable, then for any $a > 0$,

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

Proof. Let

$$I = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases}$$

be random variable. Since $X \geq 0$, $I \leq X/a$ which implies $E[I] \leq E[X]/a$. On the other hand, $E[I] = P\{I = 1\} = P\{X \geq a\}$ and that leads to the inequality above.

Theorem 4.22. Let X be a random variable with finite mean μ and variance σ^2, then for any $\epsilon > 0$,

$$P\{|X - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}$$

Proof. Let $Y = |X - \mu|^2$, then by *Theorem 4.21*,

$$P\{|X - \mu| \geq \epsilon\} = P\left\{Y \geq \epsilon^2\right\} \leq \frac{E[Y]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}$$

Example 4.23. Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50.

(a) What can be said about the probability that this week's production will exceed 75?

(b) If the variance of a week's production is known to equal 25, then what can be said about the probability that this week's production will be between 40 and 60?

Answer. *Part (a)* Let X be the number of items produced during a week with $E[X] = 50$, then by Markov's inequality (*Theorem 4.21*),

$$P\{X > 75\} = P\{P \geq 76\} \leq \frac{E[X]}{76} = \frac{25}{38}$$

Part (b) Note That

$$P\{40 \leq X \leq 60\} = P\{|X - 50| \leq 10\} = 1 - P\{|X - 50| > 10\}$$

By Chebyshev's inequality (*Theorem 4.22*),

$$P\{|X - 50| > 10\} \leq \frac{\text{Var}(X)}{10^2} = \frac{1}{4}$$

Therefore $P\{40 \leq X \leq 60\} = 3/4$.

Proposition 4.24. Let X be a random variable with finite mean μ and variance 0, then $P\{X = \mu\} = 1$.

Proof. By Chebyshev's inequality (*Theorem 4.22*),

$$\begin{aligned} P\{X \neq \mu\} &= P\left\{\bigcup_{n=1}^{\infty} |X - \mu| \geq \frac{1}{k}\right\} \\ &= \sum_{n=1}^{\infty} P\left\{|X - \mu| \geq \frac{1}{k}\right\} \\ &\leq \sum_{n=1}^{\infty} k^2 \text{Var}(X) = 0 \end{aligned}$$

which implies $P\{X = \mu\} = 1 - P\{X \neq \mu\} = 1$.

4.5.2 Weak Law of Large Numbers

Recall the problem of finding limiting behaviour of $(X_1 + X_2 + \cdots + X_n)/n$ when n tends to infinity. Below is the **weak law of large numbers**:

Theorem 4.25. Let $X_1, X_2, \cdots, X_n$ be an independent and identically distributed sequence of random variables with finite mean, then for any $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0$$

as n tends to infinity.

Proof. Assume $\text{Var}(X_i)$ is finite for all i, then note that the mean and the variance are

$$E\left[\frac{X_1+X_2+\cdots+X_n}{n}\right]=\frac{1}{n}\sum_{i=1}^{n}E[X_i]=\mu$$

and

$$\begin{aligned}\mathrm{Var}\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)&=\frac{1}{n^2}\mathrm{Var}(X_1+X_2+\cdots+X_n)\\&=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)=\frac{\sigma^2}{n}\end{aligned}$$

By applying Chebyshev's inequality (*Theorem 4.22*) to $(X_1+X_2+\cdots+X_n)/n$,

$$\begin{aligned}P\left\{\left|\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right|\geq\epsilon\right\}&\leq\frac{1}{\epsilon^2}\mathrm{Var}\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)\\&=\frac{\sigma^2}{n\epsilon^2}\to 0\end{aligned}$$

as n tends to infinity.

4.5.3 Central Limit Theorem

Proposition 4.26. Let $Z_1, Z_2, \cdots, Z_n$ be a sequence of random variables with distribution function F_{Z_n}, and Z be random variable with distribution function F_Z. Suppose $M_{Z_n}(t)\to M_Z(t)$ for all $t\in\mathbb{R}$ as n tends to infinity, then $F_{Z_n}(t)\to F_Z(t)$ for each t at which F_Z is continuous, as n tends to infinity.

Below is the **central limit theorem**:

Theorem 4.27. Let $X_1, X_2, \cdots, X_n$ be an independent and identically distributed sequence of random variables with finite mean μ and variance σ^2, then for any $a\in\mathbb{R}$,

$$P\left\{\frac{X_1+X_2+\cdots+X_n-n\mu}{\sqrt{n}\sigma}\right\}\to\Phi(a)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a}e^{-x^2/2}\,\mathrm{d}x$$

as n tends to infinity.

Proof. Assume $\mu=0$ and $\sigma^2=1$. Let $Z_n=(X_1+X_2+\cdots+X_n)/\sqrt{n}$, and Z be the standard normal random variable, then $M_Z(t)=e^{t^2/2}$ by *Example 4.17*. Note that

$$\begin{aligned}M_{Z_n}(t)&=E\left[\exp\left(t\frac{X_1+X_2+\cdots+X_n}{\sqrt{n}}\right)\right]\\&=\prod_{i=1}^{n}E[e^{tX_i/\sqrt{n}}]=\left(M_X\left(\frac{t}{\sqrt{n}}\right)\right)^n\end{aligned}$$

where $X=X_1$ for simplicity. Further let $L(t)=\log(M_X(t))$, with

$$L'(t) = \frac{M_X'(t)}{M_X(t)},\ L''(t) = \frac{M_X''(t)M_X(t) - M_X'(t)^2}{M_X(t)^2}$$

Note that when $t = 0$,

$$L'(0) = \frac{M_X'(0)}{M_X(0)} = E[X] = 1$$

and

$$L''(0) = \frac{M_X''(0)M_X(0) - M_X'(0)^2}{M_X(0)^2} = E[X^2] = \mathrm{Var}(X) + E[X]^2 = 1$$

Hence

$$\begin{aligned}\lim_{n\to\infty} nL\left(\frac{t}{\sqrt{n}}\right) &= \lim_{n\to\infty} \frac{L(t/\sqrt{n})}{(1/\sqrt{n})^2} = \lim_{x\to 0} \frac{L(tx)}{x^2} \\ &= \lim_{x\to 0} \frac{t^2 L''(tx)}{2} = \frac{t^2}{2} L''(0) = \frac{t^2}{2}\end{aligned}$$

In other words,

$$n \log\left(M_X\left(\frac{t}{\sqrt{n}}\right)\right) \to \frac{t^2}{2}$$

as n tends to infinity implies $M_{Z_n}(t) \to e^{t^2/2}$ as n tends to infinity. Generally,

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{1}{\sqrt{n}}\left(\frac{X_1 - \mu}{\sigma} + \frac{X_2 - \mu}{\sigma} + \cdots + \frac{X_n - \mu}{\sigma}\right)$$

then by taking $\tilde{X}_i = (X_i - \mu)/\sigma$ such that it has mean 0 and variance 1 finishes the proof.

In other words, the distribution of

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{\mathrm{Var}(X_1 + X_2 + \cdots + X_n)}}$$

converges to the standard normal distribution as n tends to infinity.

Example 4.28. If 10 fair dice are rolled, find the approximate probability that the sum obtained is between 30 and 40.

Answer. Let X_i be the value obtained in the i-th roll where $i = 1, 2, \cdots, 10$. Note that $E[X_i] = (1+2+3+4+5+6)/6 = 7/2$, $E[X_i^2] = (1^2+2^2+3^2+4^2+5^2+6^2)/6 = 91/6$ and $\mathrm{Var}(X_i) = 35/12$. After continuity correction, the probability required is

$$
\begin{aligned}
&P\{29.5 \le X_1 + X_2 + \cdots + X_{10} \le 40.5\} \\
=&P\left\{\frac{29.5-35}{\sqrt{350/12}} \le \frac{X_1 + X_2 + \cdots + X_{10} - 35}{\sqrt{350/12}} \le \frac{40.5-35}{\sqrt{350/12}}\right\} \\
\approx&P\{-1.02 \le Z \le 1.02\} \\
=&2\Phi(1.02) - 1 = 0.6922
\end{aligned}
$$

4.5.4 Strong Law of Large Numbers

Proposition 4.29. Let X be a nonnegative random variable with finite $E[X]$, then $P\{X < \infty\} = 1$.

Proof. By Markov's inequality,

$$P\{X = \infty\} \le P\{X \ge n\} \le \frac{E[X]}{n} \to 0$$

as n tends to infinity.

Similar to the weak law of large numbers, the **strong law of large numbers** provide a stronger estimate.

Theorem 4.30. Let $X_1, X_2, \cdots, X_n$ be an independent and identically distributed sequence of random variables with finite mean μ, then

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu$$

as n tends to infinity.

Proof. Assume $E[X_i^4] = K < \infty$, and without loss of generality, $\mu = 0$. Let $S_n = X_1 + X_2 + \cdots + X_n$ to estimate $E[S_n^4]$. First, expand $(X_1 + X_2 + \cdots + X_n)^4$ in terms of X_i^4, $X_i^3 X_j$, $X_i^2 X_j^2$, $X_i^2 X_j X_k$ and $X_i X_j X_k X_l$ where i, j, k, l are distinct. Note that

$$E[X_i^3 X_j] = E[X_i^2 X_j X_k] = E[X_i X_j X_k X_l] = 0$$

Hence

$$E[S_n^4] = nE[X_i^4] + 6\binom{n}{2} E[X_i^2] E[X_j^2]$$

Using a simple inequality $E[X^2]^2 \le E[X^4]$ from applying X^2 to definition of variance,

$$E[S_n^4] \le \left(n + 6\binom{n}{2}\right) K = (3n^2 - 2n)K \le 3n^2 K$$

implies

$$E\left[\frac{S_n^4}{n^4}\right] \leq \frac{3K}{n^2}$$

and

$$E\left[\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4\right] = \sum_{n=1}^{\infty} E\left[\left(\frac{S_n}{n}\right)^4\right] \leq \sum_{n=1}^{\infty}\frac{3K}{n^2} < \infty$$

Finally, let

$$X = \sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4$$

since $E[X]$ is finite, by *Proposition 4.29*,

$$P\left\{\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4 < \infty\right\} = 1$$

which implies

$$P\left\{\lim_{n\to\infty}\frac{S_n}{n} = 0\right\} = 1$$

and by *Proposition 4.24*,

$$\frac{S_n}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n} \to 0$$

when n tends to infinity. If $\mu \neq 0$, let $\tilde{X_n} = X_n - \mu$ and apply the process above gives

$$\frac{\tilde{X_1} + \tilde{X_2} + \cdots + \tilde{X_n}}{n} \to 0 \Leftrightarrow \frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu$$

In other words, the probability

$$P\left\{\lim_{n\to\infty}\frac{X_1 + X_2 + \cdots + X_n}{n} = \mu\right\} = 1$$

References

The following are the references of the context of this document:

(a) Course material from various professors associated to *MATH3280: Introductory Probability*

(b) S. Ross, *A First Course in Probability*, Pearson (8th Edition), 2009