MA2116 PROBABILITY

Q

Abstract. This set of notes was written during the author's first year of undergraduate study. It is not guaranteed that claims made in this set of notes are correct, and, if there are any mistakes, they will almost surely be made by the author.

Contents

1. Notations and conventions

In the first chapter, we introduce some logical symbols that will be used frequently in this set of notes.

First, we use $\mathbb{N}$ to denote the set of natural numbers, $\mathbb{Z}$ the set of integers, $\mathbb{Q}$ the set of rational numbers, and $\mathbb{R}$ the set of real numbers. In this set of notes, we adopt the convention that 0 is not a natural number. For any of the above symbols, we use X^+ to denote the set of positive elements in X and use X_0^+ to the set of non-negative elements in X. For example, $\mathbb{Z}_0^+$ should be understood as the set of non-negative integers.

Then, we have logical connectives: $\neg$ (not), $\wedge$ (and), and $\vee$ (or). For instance, we can write if A and B are both true, then at least one of C and D is true as

$$A \wedge B \implies C \vee D.$$

Here, $\implies$ stands for implication. Similarly, we use $\iff$ to express "if and only if", which is sometimes abbreviated as "iff".

Furthermore, we have universal quantifier $\forall$, read as "for all", and existential quantifier $\exists$, read as "there exists". For instance, we can write for all $x \in \mathbb{Z}^+$, there exists a $y \in \mathbb{Z}^+$ such that $y > x$ as

$$(\forall x \in \mathbb{Z}^+)(\exists y \in \mathbb{Z}^+)(y > x).$$

Less frequently used notations will be introduced whenever they are needed.

Here are some useful definitions/results:

1) (Equality of sets) $A = B \iff (\forall x)(x \in A \iff x \in B)$.

2) (Distributive law for logical connectives)

$$(A \wedge B) \vee C = (A \vee C) \wedge (A \vee B).$$

2. Combinatorial Analysis

Theorem 2.1. *(The basic principle of counting) Suppose that two experiments are to be performed. If*

- *Experiment 1 can result in any one of m possible outcomes; and*
- *Experiment 2 can result in any one of n possible outcomes;*

then together there are mn possible outcomes of the two experiments.

Theorem 2.2. *(The generalised basic principle of counting) Suppose that r experiments are to be performed. If*

- *Experiment i can result in any one of n_i possible outcomes for $1 \leq i \leq r, i \in \mathbb{Z}^+$;*

then together there are $\prod_{i=1}^{r} n_i$ possible outcomes of the r experiments.

Theorem 2.3. *(Permutations) Suppose there are n distinct objects, then the total number of different arrangements is*

$$n(n-1)(n-2)\dots(3)(2)(1) = n!$$

with the convention that $0! = 1$.

Theorem 2.4. *(Permutations with repetition) For n objects in r groups of n_i elements for $1 \leq i \leq r$, if all objects in the same group are identical and objects in different groups are distinct, then there are*

$$\frac{n!}{\prod_{i=1}^{r}(n_i!)}$$

different permutations of the n objects.

Remark 2.5. We first permute all objects as if they are all distinct, which can be done in $n!$ ways. We then eliminate the repetitions through dividing $n!$ by $\prod_{i=1}^{r}(n_i!)$, where $n_i!$ is the number of permutations of the i-th identical objects if they are viewed distinct.

Theorem 2.6. *(Circular arrangement) For n people sitting in a circle, there are*

$$\frac{n!}{n} = (n-1)!$$

possible arrangements.

Theorem 2.7. *If there are n distinct objects, of which we choose a group of r items ($1 \leq r \leq n$), then the number of possible groups is given by*

$$\begin{aligned} {}_nC_r = \binom{n}{m} &= \frac{\prod_{i=0}^{r-1} i}{r!} \\ &= \frac{\prod_{i=0}^{r-1}(n-i)}{r!} \times \frac{(n-r)!}{(n-r)!} \\ &= \frac{n!}{r!(n-r)!} \end{aligned}$$

Remark 2.8. The following results are taken for granted:

(i) For $1 \leq r \leq n$, we have

$$\binom{n}{r} = \binom{n}{n-r}$$

(ii) We have

$$\binom{n}{0} = \binom{n}{n} = 1$$

(iii) When ($n \in \mathbb{Z}_0^+$) and ($r < 0$ or $r > n$), we take

$$\binom{n}{r} = 0,$$

where $\mathbb{Z}_0^+$ is the set of all nonnegative integers.

Theorem 2.9. *For $1 \leq r \leq n$, we have*

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}$$

Theorem 2.10. *(The binomial theorem) Let $n \in \mathbb{Z}_0^+$, then*

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

Definition 2.11. (Multinomial coefficients) The number of ways to divide n objects into r distinct groups of size n_i for $1 \leq i \leq r$ such that $\sum_{i=1}^{r} n_i = n$ is given by

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\ldots-n_{r-1}}{n_r}$$

The above expression is equivalent to

$$\frac{n!}{\prod_{i=1}^{r} n_i!},$$

which is denoted by

$$\binom{n}{n_1, n_2, \ldots, n_r}$$

Theorem 2.12. *(The multinomial theorem) Let n be a nonnegative integer, then*

$$\left(\sum_{i=1}^{r} x_i\right)^n = \sum_{\sum_{i=1}^{r} n_i = n} \binom{n}{n_1, n_2, \ldots, n_r} \prod_{i=1}^{r} x_i^{n_i}$$

Theorem 2.13. *(Balls and urns model) The number of ways to distribute n identical balls to r distinct urns is given by*

$$\binom{n+r-1}{r-1}.$$

Remark 2.14. To illustrate the above theorem, we consider a special case of $n = 5$ and $r = 3$. Suppose we have 5 identical stones and $(3-1) = 2$ distinct sticks. We can arrange them as

$$\bullet \mid \bullet \ \bullet \mid \bullet \ \bullet.$$

This arrangement can be considered to be corresponding to the distribution of putting 1 ball into the first urn and 2 balls each in the remaining two urns. If we arrange stones and sticks as

$$\bullet \mid \bullet \mid \bullet \ \bullet \ \bullet,$$

then we can just put one ball in each of the first two urns and the remaining three balls into the last urn. With this idea in mind, we can construct a bijection between the ways we put balls into the urns and the number of such permutations, which affirms the validity of the above theorem.

Example 2.15. Find the number of non-negative integer solutions of the equation

$$x + y + z = 10.$$

Answer. This is equivalent to finding the number of ways to distribute 10 identical balls into 3 distinct urns (x, y, z). Hence, the number of non-negative integer solutions is given by $\binom{12}{2} = 66$.

3. Axioms of Probability

We assume that the readers have some knowledge of sets.

Definition 3.1. (Sample space) The sample space is the set of all possible outcomes of an experiment, usually denoted by S.

Definition 3.2. (Event) Any subset A of the sample space is an event.

Example 3.3. Suppose we flip two coins. The sample space S of the experiment is

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

Remark 3.4. In the above example, we should take (H, H) as an outcome and $\{(H, H)\}$ as an event. In other words, we can consider an event to be a collection of outcome(s).

Definition 3.5. (Set operations) Let A and B be subsets of S.

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

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

Difference: $A \backslash B = \{x \in S : x \in A \text{ and } x \notin B\}$

Complement: $A^c = \{x \in S : x \notin A\}$

Remark 3.6. In this set of notes, $A \cap B$ is sometimes abbreviated as AB.

Remark 3.7. Some textbooks use $A - B$ to denote the set difference. Since $A - B$ has different meanings under different contexts, we choose $A \backslash B$ in this set of notes.

Proposition 3.8. *We have the following results on the operations of sets:*

- *Commutative laws*

  $E \cap F = F \cap E$

  $E \cup F = F \cup E$

- *Associative laws*

  $(E \cap F) \cap G = E \cap (F \cap G)$

  $(E \cup F) \cup G = E \cup (F \cup G)$

- *Distributive laws*

  $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$

  $(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$

- *De Morgan's laws*

  $(\bigcup_{i=1}^{n} E_i)^c = \bigcap_{i=1}^{n} E_i^c$

  $(\bigcap_{i=1}^{n} E_i)^c = \bigcup_{i=1}^{n} E_i^c$

Proof. The proof for commutative laws and associative laws should be obvious. Hence, we only present the proof for distributive laws and De Morgan's laws here. Let x be an arbitrary element in S.

We first prove distributive laws:

$$\begin{aligned} x \in (E \cup F) \cap G &\iff x \in (E \cup F) \wedge x \in G \\ &\iff (x \in E \vee x \in F) \wedge x \in G \\ &\iff (x \in E \wedge x \in G) \vee (x \in F \wedge x \in G) \\ &\iff (x \in E \cap G) \vee (x \in F \cap G) \\ &\iff x \in (E \cap G) \cup (F \cap G). \end{aligned}$$

Since this true for all x, we conclude that $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$. Similarly, we can prove the other distributive law.

We proceed to prove the De Morgan's laws, which can be done by induction.

When $n = 2$, we have

$$\begin{aligned} x \in (E_1 \cup E_2)^c &\iff x \in S \wedge x \notin (E_1 \cup E_2) \\ &\iff x \in S \wedge (x \notin E_1 \wedge x \notin E_2) \\ &\iff (x \in S \wedge x \notin E_1) \wedge (x \in S \wedge x \notin E_2) \\ &\iff (x \in E_1^c) \wedge (x \in E_2^c) \\ &\iff x \in E_1^c \cap E_2^c. \end{aligned}$$

We assume the case $n = k$ is true, and now prove the case of $n = k + 1$:

$$\begin{aligned} x \in \left(\bigcup_{i=1}^{k+1} E_i\right)^c &\iff x \in \left(E_{k+1} \cup \left(\bigcup_{i=1}^{k} E_i\right)\right)^c \\ &\iff x \in \left(E_{k+1}^c \cap \left(\bigcup_{i=1}^{k} E_i\right)^c\right) \\ &\iff x \in \left(E_{k+1}^c \cap \left(\bigcap_{i=1}^{k} E_i^c\right)\right) \\ &\iff x \in \bigcap_{i=1}^{k+1} E_i^c. \end{aligned}$$

We thus prove the case of $n = k + 1$. Therefore, by the principle of mathematical induction, we conclude that $(\bigcup_{i=1}^{n} E_i)^c = \bigcap_{i=1}^{n} E_i^c$. By adopting a similar approach, we can prove that $(\bigcap_{i=1}^{n} E_i)^c = \bigcup_{i=1}^{n} E_i^c$. □

With these ideas in mind, we now try to make sense of what probability is. There are a few ways to define probability:

(1) Classical approach

Assume all the sample points are equally likely events. We can then define $P(E)$ as

$$P(E) = \frac{|E|}{|S|},$$

where $|\cdot|$ refers to the number of elements in the set $\cdot$.

(2) Relative frequency approach

This is largely based on the experiments:

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

where $n(E)$ is the number of times in n repetitions of the experiment that E occurs.

(3) Subjective approach

In this approach, probability is considered as a measure of belief.

Remark 3.9. In MA2116, we mainly consider the classical approach and ignore the other two.

Definition 3.10. (Mutually exclusive events) Two events A, B are said to be mutually exclusive if $A \cap B = \emptyset$.

Remark 3.11. If an outcome of an event is observed, we say that this specific event happens.

Remark 3.12. $A \cap B$ contains all the outcomes in both A and B. With this regard, $A \cap B = \emptyset$ implies that A and B contain no common outcomes, so it is impossible for them to occur simultaneously. We thus call A and B mutually exclusive events.

Definition 3.13. (Axioms of probability) Probability, denoted by P, is a function on the collection of events satisfying

(1) For any event E, $P(E) \geq 0$.

(2) Let S be the sample space, then $P(S) = 1$.

(3) For any sequence of pairwise mutually exclusive events $\{E_i\}$, we have

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

Remark 3.14. We call $P(E)$ the probability of the event E.

Remark 3.15. In the lecture notes, the axiom (i) states $(\forall E \in F)(0 \leq P(E) \leq 1)$, where F is the event space, i.e. the collection of all the events. We choose to state a weaker version of axiom (i) and show that $P(E) \leq 1$ later as a proposition.

Proposition 3.16. $P(\emptyset) = 0$.

Proof. Take $E_1 = S$ and $E_i = \emptyset$ for $i \geq 2, i \in \mathbb{Z}^+$. It is easy to verify that $\{E_i\}$ forms a sequence of pairwise mutually exclusive events.

By Axiom (iii), we have

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = P(S) + \sum_{i=2}^{\infty} P(E_i).$$

Since $(\bigcup_{i=1}^{\infty} E_i) = S$, we have

$$P(S) = P(S) + \sum_{i=2}^{\infty} P(E_i)$$

$$\sum_{i=2}^{\infty} P(E_i) = 0.$$

By axiom (i), we have $(\forall i \in \mathbb{Z}^+)(P(E_i) \geq 0)$, so the above equation implies that $P(E_i) = 0$ for all $i \geq 2$. We therefore conclude that $P(\emptyset) = 0$. □

Remark 3.17. This coincides with our intuition that the empty set as an event implies nothing is observed, which cannot be the case, i.e., the probability is 0.

Proposition 3.18. *For any finite collection of pairwise mutually exclusive events* $\{E_i\}, 1 \leq i \leq n$*, we have*

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

Proof. Define a new sequence $\{T_i\}$ by setting $T_i = E_i$ for $1 \leq i \leq n$ and $T_i = \emptyset$ for $i > n$. It can be verified that $\{T_i\}$ is a sequence of pairwise mutually exclusive events. We thus have

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) &= P\left(\bigcup_{i=1}^{n} T_i\right) + P\left(\bigcup_{i=n+1}^{\infty} T_i\right) \\ &= P\left(\bigcup_{i=1}^{\infty} T_i\right) \\ &= \sum_{i=1}^{\infty} P(T_i) \\ &= \sum_{i=1}^{n} P(T_i) + \sum_{i=n+1}^{\infty} P(\emptyset) \\ &= \sum_{i=1}^{n} P(E_i) \end{aligned}$$

□

Proposition 3.19. *Let E be an event, then*

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

Proof. By the definition of set union and set complement, we have

$$E \cup E = S^c, \quad E \cap E^c = \emptyset.$$

Hence, we have

$$P(E \cup E^c) = P(E) + P(E^c)$$
$$P(S) = P(E) + P(E^c)$$
$$P(E^c) = 1 - P(E)$$

□

Corollary 3.20. *If $A \subseteq S$, then $P(A) \leq 1$*

Proof. By the axiom (i) and the above proposition, we have

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

□

Proposition 3.21. *If $A \subseteq B$, then $P(A) \leq P(B)$.*

Proof. We can express B as $B = (B \cap A) \cup (B \cap A^c)$. Since A and A^c are mutually exclusive, we must have that $B \cap A$ and $B \cap A^c$ are mutually exclusive events. Hence, we have

$$\begin{aligned} P(B) &= P[(B \cap A) \cup (B \cap A^c)] \\ &= P(B \cap A) + P(B \cap A^c) \\ &= P(A) + P(B \cap A^c) \\ &\geq P(A) \end{aligned}$$

□

Proposition 3.22. $P(A\backslash B) = P(A) - P(A \cap B)$

Proof. By the definition of set difference and set intersection, we have $A = (A\backslash B) \cup (A \cap B)$, where $A\backslash B$ and $A \cap B$ are mutually exclusive events given that $\in$ is well-defined. We thus have

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

□

Proposition 3.23. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof. We can decompose $A \cup B$ as $(A\backslash B) \cup (A \cap B) \cup (B\backslash A)$, which are pairwise mutually exclusive.

Hence, we have

$$\begin{aligned} P(A \cup B) &= P(A\backslash B) + P(A \cap B) + P(B\backslash A) \\ &= P(A\backslash B) + P(A \cap B) + P(B\backslash A) \\ &= P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

□

Remark 3.24. When we try to prove certain properties, we can make use of the fact that an event is a subset of the sample space. This means, we can consider using set operations to decompose an event into smaller disjoint sub-events.

Remark 3.25. Proposition 3.23 is the base case of the following theorem.

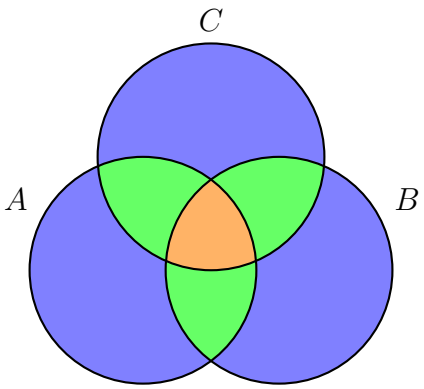
Theorem 3.26. (Inclusion/Exclusion Principle or PIE) *Let* $\{E_i\}$ *be a collection of events for* $1 \leq i \leq n, i \in \mathbb{Z}^+$. *We have*

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \leq i_1 < \ldots < i_r \leq n} P(E_{i_1} \cap \ldots \cap E_{i_r}).$$

Example 3.27. When $n = 3$, the formula becomes

$$\begin{aligned} P\left(\bigcup_{i=1}^{3} E_i\right) &= \sum_{r=1}^{3}(-1)^{r+1} \sum_{1\le i_1<\ldots<i_r\le n} P(E_{i_1} \cap \ldots \cap E_{i_r}) \\ &= \sum_{i=1}^{3} P(E_i) - \sum_{1\le i_1<i_2\le 3} P(E_{i_1} \cap E_{i_2}) + P\left(\bigcup_{i=1}^{3} E_i\right) \\ &= \sum_{i=1}^{3} P(E_i) - [P(E_1 \cap E_2) + P(E_2 \cap E_3) + P(E_1 \cap E_3)] + P\left(\bigcap_{i=1}^{3} E_i\right) \end{aligned}$$

Remark 3.28. To make sense of this example, we can plot a Venn diagram to visualise it.



In our approach, when we calculate $\sum_{i=1}^{3} P(E_i)$, we double count the green regions and triple count the orange region. To correct this, we then subtract $\sum_{1\le i_1<i_2\le 3} P(E_{i_1} \cap E_{i_2})$ from the expression to remove the double counting. However, this means that we count the orange region trice again, so it has not been counted overall. We therefore add in the last term $P\left(\bigcap_{i=1}^{3} E_i\right)$ and get the desired probability.

We now give a formal proof of PIE.

Proof. Let P_n be the statement "$P\left(\bigcup_{i=1}^{n} E_i\right)$ can be computed by PIE given $\{E_i\}$ is a sequence of events for $1 \le i \le n, i \in \mathbb{Z}^+$"

When $n = 1$, we have the trivial case $P(E_1) = P(E_1)$.

When $n = 2$, we have proposition 3.23.

Suppose P_k is true. When $n = k + 1$, we have

$$\begin{aligned} LHS = P\left(\bigcup_{i=1}^{k+1} E_i\right) &= P\left[\left(\bigcup_{i=1}^{k} E_i\right) \cup E_{k+1}\right] \\ &= P\left(\bigcup_{i=1}^{k} E_i\right) + P(E_{k+1}) - P\left[\left(\bigcup_{i=1}^{k} E_i\right) \cap E_{k+1}\right] \\ &= P\left(\bigcup_{i=1}^{k} E_i\right) + P(E_{k+1}) - P\left[\bigcup_{i=1}^{k} (E_i \cap E_{k+1})\right] \\ &= \left[\sum_{r=1}^{k} (-1)^{r+1} \sum_{1 \le i_1 < \ldots < i_r \le k} P(E_{i_1} \cap \ldots \cap E_{i_r})\right] + P(E_{k+1}) + P\left[\bigcup_{i=1}^{k} (E_i \cap E_{k+1})\right] \end{aligned}$$

Since $P\left[\bigcup_{i=1}^{k} (E_i \cap E_{k+1})\right]$ also satisfies the form of our inductive hypothesis, we have

$$\begin{aligned} P\left[\bigcup_{i=1}^{k} (E_i \cap E_{k+1})\right] &= \sum_{r=1}^{k} (-1)^{r+1} \sum_{1 \le i_1 < \ldots < i_r \le k} P[(E_{k+1} \cap E_{i_1}) \cap \ldots \cap (E_{k+1} \cap E_{i_r})] \\ &= \sum_{r=1}^{k} (-1)^{r+1} \sum_{1 \le i_1 < \ldots < i_r \le k} P(E_{k+1} \cap E_{i_1} \cap \ldots \cap E_{i_r}). \end{aligned}$$

Consequently, we have

$$\begin{aligned} LHS &= P\left(\bigcup_{i=1}^{k} E_i\right) + P(E_{k+1}) - P\left[\bigcup_{i=1}^{k} (E_i \cap E_{k+1})\right] \\ &= \sum_{r=1}^{k+1} (-1)^{r+1} \sum_{1 \le i_1 < \ldots < i_r \le k+1} P(E_{i_1} \cap \ldots \cap E_{i_r}) \\ &= RHS. \end{aligned}$$

By the principle of mathematical induction, we conclude our proof. □

Proposition 3.29. *When all outcomes in a finite sample space are equally likely to occur, if the event A contains $|A|$ outcomes, and the sample space S contains $|S|$ outcomes, then*

$$P(A) = \frac{|A|}{|S|}$$

Proof. Let $\{s_i\}$ be an event with only one outcome s_i for $1 \leq i \leq n$. We have

$$\begin{aligned} P(S) &= P\left(\bigcup_{i=1}^{n}\{s_i\}\right) \\ 1 &= \sum_{i=1}^{n} P(\{s_i\}) \\ 1 &= nP(\{s_i\}) \\ P(\{s_i\}) &= \frac{1}{n} \end{aligned}$$

For any event A of a outcomes, we can express A as the union of the corresponding $\{s_{p_i}\}$ for some $1 \leq p_i \leq n, 1 \leq i \leq a$, which are pairwise mutually exclusive:

$$\begin{aligned} P(A) &= P\left(\bigcup_{i=1}^{a}\{s_{p_i}\}\right) \\ &= \sum_{i=1}^{a} P(\{s_{i_i}\}) \\ &= \frac{a}{n} \\ &= \frac{|A|}{|S|}. \end{aligned}$$

□

Definition 3.30. (Increasing sequence) A sequence of events $\{E_n\}, n \geq 1$ is said to be an increasing sequence if

$$E_1 \subseteq E_2 \subseteq \ldots \subseteq E_n \subseteq E_{n+1} \subseteq \ldots,$$

whereas it is said to be a decreasing sequence if

$$E_1 \supseteq E_2 \supseteq \ldots \supseteq E_n \supseteq E_{n+1} \supseteq \ldots.$$

Definition 3.31. If $\{E_n\}, n \geq 1$ is an increasing sequence of events, then we define a new event, denoted by $\lim_{n\to\infty} E_n$ as

$$\lim_{n\to\infty} E_n := \lim_{n\to\infty} \bigcup_{i=1}^{n} E_i = \bigcup_{i=1}^{\infty} E_i.$$

Similarly, if $\{E_n\}, n \geq 1$ is a decreasing sequence of events, then we define a new event, denoted by $\lim_{n\to\infty} E_n$ as

$$\lim_{n\to\infty} E_n := \lim_{n\to\infty} \bigcap_{i=1}^{n} E_i = \bigcap_{i=1}^{\infty} E_i.$$

Remark 3.32. We need to be clear that $\lim_{n\to\infty} E_n$ is just an event, and we can use any letter we want to denote it. Do not be confused by the "$\lim_{n\to\infty}$" symbol here; it is just a notation. That is why we need the following theorem.

Proposition 3.33. *If $\{E_n\}, n \geq 1$ is either an increasing sequence or a decreasing of events, then*

$$P\left(\lim_{n\to\infty} E_n\right) = \lim_{n\to\infty} P(E_n).$$

Proof. We first present the proof for the case of $\{E_n\}$ being an increasing sequence. By the definition of $\lim_{n\to\infty} E_n$, it is equivalent to proving

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

Noticing that the LHS is the probability of a union of events, we then consider using the probability axiom. To do so, we first need to define a new sequence of pairwise disjoint events $\{F_n\}_{n\in\mathbb{Z}^+}$ satisfying

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

Naturally, since $\{E_i\}$ is an increasing sequence, we can define $\{F_i\}$ by considering

$$\begin{cases} F_1 := E_1 \\ F_n := E_n \cap \left(\bigcup_{i=1}^{n-1} E_i\right)^c = E_n \cap E_{n-1}^c, \quad \text{for } n \geq 2 \end{cases}.$$

Geometrically, F_n represents the difference between E_n and E_{n-1}, so $\{F_n\}$ must be pairwise disjoint. Moreover, we can verify that $\bigcup_{i=1}^{n} E_i = \bigcup_{i=1}^{n} F_i$ for all $n \in \mathbb{Z}^+$. With these two

conditions, we have

$$\begin{aligned} LHS = P\left(\lim_{n\to\infty} E_n\right) = P\left(\bigcup_{i=1}^{\infty} E_i\right) &= P\left(\bigcup_{i=1}^{\infty} F_i\right) \\ &= \sum_{i=1}^{\infty} P(F_i) && \text{(By the axiom 3)} \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P(F_i) && \text{(By definition)} \\ &= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} F_i\right) && \text{(By proposition 3.18)} \\ &= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} E_i\right) && \text{(By the second condition)} \\ &= \lim_{n\to\infty} P(E_n) && (\{E_n\} \text{ is increasing}) \end{aligned}$$

We therefore conclude our proof for the case of $\{E_n\}$ being increasing.

For the case of $\{E_n\}$ being decreasing, we have $\{E_n^c\}$ is increasing. With the above result, we have

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

By De Morgan's law, we have

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

which, according to proposition 3.19, gives

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

Rearranging the above equation, we get the desired result. □

Remark 3.34. Readers may find results introduced in this chapter familiar. In fact, the main difficulty of this chapter is not the complexity of the propositions presented, but how to derive these results by strict logical reasoning.

4. Conditional Probability and Independence

Definition 4.1. (Conditional probability) Let E and F be two events. Suppose that $P(F) > 0$, the conditional probability of E given F is defined as

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

Remark 4.2. By rearranging the equation, we will have $P(E \cap F) = P(E|F)P(F)$, which allows us to compute $P(E \cap F)$ based on the conditional probabilities. It is also to be noted that at this point, we are unsure about whether conditional probability follows the axioms of probability, i.e., whether conditional probabilities are indeed probabilities.

Theorem 4.3. (General Multiplication Rule) *Let $\{A_i\}_{i\in\mathbb{N}}$ be a set of events. Then we have*

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)\prod_{i=2}^{n} P\left(A_i \middle| \bigcap_{r=1}^{i-1} A_i\right).$$

Example 4.4. Let A_1, A_2, A_3 be a set of events. Then the above equation gives us

$$P\left(\bigcap_{i=1}^{3} A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2).$$

Proof. We prove this by mathematical induction. The base case can be directly derived from the definition of conditional probability. Suppose the theorem holds for $n = k$. We now prove the case of $n = k + 1$:

$$\begin{aligned}
LHS = P\left(\bigcap_{i=1}^{k+1} A_i\right) &= P\left(A_{k+1} \cap \left(\bigcap_{i=1}^{k} A_i\right)\right) \\
&= P\left(A_{k+1} \middle| \bigcap_{r=1}^{k} A_i\right) P\left(\bigcap_{i=1}^{k} A_i\right) \\
&= P\left(A_{k+1} \middle| \bigcap_{r=1}^{k} A_i\right) \left[P(A_1)\prod_{i=2}^{k} P\left(A_i \middle| \bigcap_{r=1}^{k-1} A_i\right)\right] \\
&= P(A_1)\prod_{i=2}^{k+1} P\left(A_i \middle| \bigcap_{r=1}^{k} A_i\right) = RHS
\end{aligned}$$

We therefore conclude our proof. □

Theorem 4.5. (Bayes' Formula) *Let A and B be any two events, then*

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

Proof. We use Venn's diagram to dissect B as a set.

$$\begin{aligned} P(B) &= P\left(B \cap (A \cup A^c)\right) \\ &= P\left((B \cap A) \cup (B \cap A^c)\right) \\ &= P\left(B \cap A\right) + P\left(B \cap A^c\right) - P(B \cap A \cap B \cap A^c) \\ &= P\left(B \cap A\right) + P\left(B \cap A^c\right) \\ &= P\left(B|A\right) P(A) + P\left(B|A^c\right) P(A^c) \end{aligned}$$

We thus conclude our proof. □

Definition 4.6. (Partition) We say that a finite collection of the subsets of the sample space S, $\{A_i\}_{i\in I}$, is a partition of S if

(1) (mutually exclusive) $(\forall i \in I)(\forall j \in I)(A_i \cap A_j = \emptyset)$

(2) (collectively exhaustive) $\bigcup_{i\in I} A_i = S$.

Remark 4.7. Usually, the index set I is in the form of $\{1, 2, \ldots, n\}$. If an index set contains integers from 1 to n, we may use the notation I_n to denote this particular index set, but since the value of n is not important here, we just drop the subscript and use I instead.

Proposition 4.8. *(Bayes' First Formula) Suppose the finite collection of events $\{A_i\}_{i\in I}$ is a partition of the sample space. Assume further that $P(A_i) > 0$ for $1 \leq i \leq n$. Let B be any event. We have*

$$P(B) = \sum_{i\in I} P(B|A_i)P(A_i).$$

Proof. Since $\{A_i\}$ is the partition of the sample space, we have

$$P(B) = P(B \cap S) = P\left[B \cap \left(\bigcup_{i\in I} A_i\right)\right] = P\left[\bigcup_{i\in I}(B \cap A_i)\right].$$

Furthermore, since $\{A_i\}$ are mutually exclusive, we have $(B \cap A_i)$ are pairwise disjoint for all $i \in I$. Therefore, by the probability axiom and the definition of conditional probability, we

conclude that

$$P(B) = P\left[\bigcup_{i\in I}(B \cap A_i)\right] = \sum_{i\in I} P(B \cap A_i) = \sum_{i\in I} P(B|A_i)P(A_i).$$

□

Remark 4.9. This can be seen as a generalisation of the Bayes' formula (theorem 4.5).

Proposition 4.10. *(Bayes' Second Formula) Suppose the events $A_1, \ldots, A_n$ partition the sample space, and assume that $P(A_i) > 0$ for all $1 \leq i \leq n$.Let B by any event, then for any $1 \leq i \leq n$, we have*

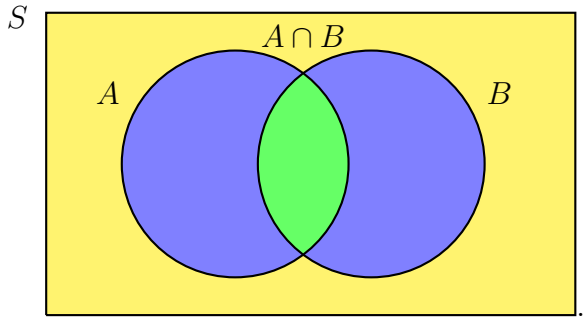
$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{r=1}^{n} P(B|A_r)P(A_r)}.$$

Proof. It is a direct implication of Bayes' First Formula and the definition of conditional probability:

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i \cap B)}{P(B)} \\ &= \frac{P(B|A_i)P(A_i)}{\sum_{r=1}^{n} P(B|A_r)P(A_r)}. \end{aligned}$$

□

Remark 4.11. We consider the following Venn's diagram:



Since we assume that all outcomes are equally likely, we may view $P(A)$ as the proportion of S which A occupies, so the condition probability $P(A|B)$ can thus be understood as the proportion of B which A occupies (or more specifically $A \cap B$ occupies). With this understanding, the Venn's diagram above should provide a good illustration of Bayes' second formula.

Definition 4.12. (Odds) The odds of an event A is defined by

$$odds = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}.$$

Remark 4.13. When we are unable to calculate the probability due to a lack of information, sometimes, we can calculate odds to assess whether A is likely to happen.

Definition 4.14. (Independent events) Two events A and B are said to be independent if

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

If two events are not independent, then they are said to be dependent events.

Remark 4.15. The intuitive idea of independent events is that the occurrence of one event does not affect the occurrence of the other event. Therefore, we are expected to have

$$P(A|B) = P(A),$$

rearranging which gives us the above definition.

Proposition 4.16. *If A and B are independent, then so are*

(i) A and B^c;

(ii) A^c and B;

(iii) A^c and B^c.

Proof. By symmetry, *(ii)* and *(iii)* will be the direct implication of *(i)*, so we then just present the proof for *(i)* here.

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

□

Remark 4.17. Even if A is independent of B and C respectively, it may NOT be true that A is independent of $B \cap C$.

Example 4.18. Two fair dice are thrown. Let A be the event that the sum of the dice is 7, B the event that the first die is 4, C that the second die is 3. It is easy to get that

$$P(A) = P(B) = P(C) = \frac{1}{6}; P(A \cap B) = P(A \cap C) = P(A \cap B \cap C) = \frac{1}{36},$$

which suggests that $P(A \cap B \cap C) \neq P(A)P(B)P(C)$. The key thing here is that, once A and B are observed simultaneously, C is naturally observed.

Definition 4.19. (Independent events) Three events A, B, C are said to be independent if they are pair-wise independent and

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

Definition 4.20. (Independent events) Let $\{A_i\}$ be a finite collection of events. All events in $\{A_i\}$ are said to be independent if for every subcollection of events $S := \{A_{i_j}\}$, we have

$$P\left(\bigcap_{A_{i_j} \in S} A_{i_j}\right) = \prod_{A_{i_j} \in S} P(A_{i_j}).$$

Remark 4.21. Suppose we have a collection of events $\{A_i\}$ for $1 \leq i \leq 4$. Following the definition, to say that these four events are independent, we must have that they are pair-wise independent and for any of the three events, they are also independent. This is equivalent to saying that any subcollection S of $\{A_i\}$, say $\{A_1, A_2, A_4\}$, satisfies the above equation.

Proposition 4.22. *If A, B, and C are independent, then A is also independent of $B \cup C$ and $B \cap C$.*

Proof. We first show that $P(A \cap (B \cup C)) = P(A)P(B \cup C)$. By the principle of inclusion and exclusion, we have

$$\begin{aligned} P(A \cap (B \cup C)) &= P[(A \cap B) \cup (A \cap C)] \\ &= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \\ &= P(A)P(B) + P(A)P(C) - P(A)P(B)P(C) = P(A)P(B \cup C). \end{aligned}$$

We then prove $P(A \cap (B \cap C)) = P(A)P(B \cap C)$. Since A, B, C are independent, we have

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

We therefore conclude our proof. □

Proposition 4.23. *If A is an event with $P(A) > 0$, then the following three conditions hold:*

(i) For any event B, we have $0 \leq P(B|A) \leq 1$.

(ii) $P(S|A) = 1$.

(iii) Let $B_1, B_2, B_3, ...$ be a sequence of mutually exclusive events, then

$$P\left(\bigcup_{k=1}^{\infty} B_k|A\right) = \sum_{k=1}^{\infty} P(B_k|A).$$

Proof. (i) Since we have $\emptyset \subseteq B \cap A \subseteq A$, by proposition 3.21, we have

$$P(\emptyset) \leq P(B \cap A) \leq P(A),$$

dividing all by $P(A)$ gives us $0 \leq P(B|A) \leq 1$.

(ii) We have

$$P(S|A) = \frac{P(S \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1.$$

(iii) Since $\{B_i\}$ are pairwise mutually exclusive, we have $B_i \cap A$ and $B_j \cap A$ are mutually exclusive iff $i \neq j$. This gives us

$$\begin{aligned} P\left(\bigcup_{k=1}^{\infty} B_k|A\right) &= \frac{P\left[\left(\bigcup_{k=1}^{\infty} B_k\right) \cap A\right]}{P(A)} \\ &= \frac{P\left[\bigcup_{k=1}^{\infty}(B_k \cap A)\right]}{P(A)} \\ &= \sum_{k=1}^{\infty} \frac{P(B_k \cap A)}{P(A)} \\ &= \sum_{k=1}^{\infty} P(B_k|A). \end{aligned}$$

We thus conclude our proof. □

5. Discrete Random Variables

Definition 5.1. (Random variable) A random variable, X, is a mapping from the sample space to real numbers, i.e., $X : S \mapsto \mathbb{R}$.

Remark 5.2. Using random variables allows us to treat outcomes as if they are numbers.

Example 5.3. Toss a fair coin; the set of all possible outcomes S is {head, flower}. We may thus define a random variable by setting

$$X(s) := \begin{cases} 1 & \text{if } s = \text{head;} \\ 0 & \text{if } s = \text{flower.} \end{cases}$$

Definition 5.4. (Discrete random variable) A random variable is said to be discrete if the range of X is either finite or countably infinite.

Example 5.5. The random variable defined in example 5.3 is discrete.

Definition 5.6. (Probability mass function/pmf) If a random variable X is discrete, taking values x_1, x_2, ..., then probability mass function of X, denoted by p_X, is defined as,

$$p_X(x) = \begin{cases} P(X = x) & \text{if } x = x_1, x_2, \ldots; \\ 0 & \text{otherwise.} \end{cases}$$

Example 5.7. The pmf of the random variable defined in example 5.3 is given by

$$p_X(x) := \begin{cases} \frac{1}{2} & \text{if } x = 0 \text{ or } 1; \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 5.8. *From the definition of pmf, we have*

(*i*) p_X *is non-negative;*

(*ii*) $\sum_{i=1}^{\infty} p_X(x_i) = 1$.

Definition 5.9. (Cumulative distribution function/cdf) The cumulative distribution function of X, denoted by F_X, is defined as a function $F_X : \mathbb{R} \mapsto \mathbb{R}$, $F_X(x) = P(X \leq x)$ for all $x \in \mathbb{R}$.

Example 5.10. The cdf of the random variable defined in example 5.3 is given by

$$F_X(x) := \begin{cases} 0 & \text{if } x < 0; \\ \frac{1}{2} & \text{if } 0 \leq x < 1; \\ 1 & \text{if } 1 \leq x. \end{cases}$$

Proposition 5.11. *Let X be a random variable. The cdf of X has the following properties:*

1) F_X is non-decreasing;

2) $\lim_{b\to\infty} F_X(b) = 1$, and $\lim_{b\to-\infty} F_X(b) = 0$;

3) F_X has left limits, i.e., $\lim_{x\to b^-} F_X(x)$ exists for all $b \in \mathbb{R}$;

4) F_X is right continuous, i.e., $\lim_{x\to b^+} F_X(x) = F(b)$ for all $b \in \mathbb{R}$.

Proof. 1) Let $a < b$ be arbitrarily chosen. Since

$$F_X(b) - F_X(a) = P(X \leq b) - P(X \leq a) = P(a < X \leq b) \geq 0,$$

we conclude that $a < b \implies F_X(a) \leq F_X(b)$, which suggests that F_X is non-decreasing.

2) Define $A_n := \{X \leq n\}$ for all $n \in \mathbb{Z}^+$. It is easy to verify that $\{A_n\}$ is an increasing sequence. Therefore, we have

$$\lim_{b\to\infty} F_X(b) = \lim_{b\to\infty} P(X \leq b) = \lim_{b\to\infty} P(A_b) = P\left(\bigcup_{b=1}^{\infty} A_b\right) = P(S) = 1.$$

Similarly, by defining $B_n := \{X \leq -n\}$ for all $n \in \mathbb{Z}^+$, which makes $\{B_n\}$ a decreasing sequence, we have

$$\lim_{b\to-\infty} F_X(b) = \lim_{b\to-\infty} P(X \leq b) = \lim_{b\to-\infty} P(B_b) = P\left(\bigcap_{b=1}^{\infty} B_b\right) = P(\emptyset) = 0.$$

3) Let $\{b_n\}$ be any increasing sequence converging to b. By the result of 1), we have $P(X \leq b_n)$ is non-decreasing sequence bounded above by $P(X \leq b)$ as $b_n < b$. Therefore, by the properties of real numbers, since $\{b_n\}$ is arbitrarily chosen, we conclude that $\lim_{x\to b^-} F_X(x)$ exists.

4) Let $\{b_n\}$ be any decreasing sequence converging to b, and define $D_n := \{X \leq b_n\}$, which is clearly a decreasing sequence. By the sequence criterion of continuity (not covered in this module), since we have $\{b_n\}$ is arbitrarily chosen and

$$\lim_{n\to\infty} P(X \leq b_n) = \lim_{n\to\infty} P(D_n) = P\left(\bigcap_{n=1}^{\infty} D_n\right) = P(D_\infty) = P(X \leq b) = F_X(b),$$

we conclude that $F_X(b)$ is right continuous. □

Proposition 5.12. *We have*

1) $P(a < X \leq b) = F_X(b) - F_X(a)$*;*

2) $P(X = a) = F_X(a) - F_X(a^-)$*, where* $F_X(a^-) = \lim_{x\to a^-} F_X(x)$.

Definition 5.13. (Expectation) If X is a discrete random variable having a pmf p_X, the expectation or the expected value of X, denoted by $E(X)$ or μ_X, is defined by

$$E(X) = \sum_{x\in R(X)} x p_X(x),$$

where $R(X)$ denotes the range of X.

Example 5.14. The expected value of the random variable defined in example 5.3 is given by

$$E(X) = \sum_{x\in R(X)} x p_X(x) = \frac{1}{2}(0) + \frac{1}{2}(1) = \frac{1}{2},$$

where $R(X) = \{0, 1\}$.

Remark 5.15. The expectation of X can be seen as the weighted average of all possible values that X can take, where the probability $P(X = x)$ serves as the weightage.

Proposition 5.16. *For nonnegative integer-valued random variable* X*, we have*

$$E(X) = \sum_{k=1}^{\infty} P(X \geq k) = \sum_{k=0}^{\infty} P(X > k).$$

Proof. By the definition of expectation, we have

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} xP(X=x) \\ &= \sum_{x=1}^{\infty}\sum_{k=1}^{x} P(X=x) \\ &= \sum_{k=1}^{\infty}\sum_{x=k}^{\infty} P(X=x) \\ &= \sum_{k=1}^{\infty} P(X \geq k). \end{aligned}$$

The second equality is obtained by using the fact $P(X \geq k) = P(X > k-1)$. □

Proposition 5.17. *If X is a discrete random variable that takes values x_i, $i \geq 1$ with respective probabilities $p_X(x_i)$, then for any real value function g, we have*

$$E[g(X)] = \sum_i g(x_i)p_X(x_i) = \sum_{x \in R(X)} g(x)p_X(x).$$

Proof. Let $Y := g(X)$ taking values y_j for $j \geq 1$. By the definition of expectation, we have

$$\begin{aligned} E[g(X)] = E(Y) &= \sum_j y_j P(Y=y_j) \\ &= \sum_j y_j \left(\sum_{i:g(x_i)=y_j} P(X=x_i) \right) \\ &= \sum_j \sum_{i:g(x_i)=y_j} g(x_i)p_X(x_i) \\ &= \sum_i g(x_i)p_X(x_i), \end{aligned}$$

with the observation that $\sum_j \sum_{i:g(x_i)=y_j} = \sum_i$. This is a valid observation since the LHS is just a regrouping of elements at the RHS, where all x_i satisfying $g(x_i) = y_j$ are grouped together. □

Corollary 5.18. $E(aX+b) = aE(X) + b$.

Definition 5.19. (Moment) Let X be a discrete random variable. For all $k \geq 1$, $E(X^k)$ is said to be the k-th moment of X.

Definition 5.20. (Central moment) Let X be a discrete random variable, $\mu = E(X)$, and $g(x) = (x - \mu)^k$. We call $E[g(X)] = E[(X - \mu)^k]$ the k-th central moment of X.

Definition 5.21. (Indicator) For any event $A \subseteq S$, we define I_A to be the indicator of A as

$$I_A(s) = \begin{cases} 1 & \text{if } s \in A; \\ 0 & \text{if } s \in S \backslash A. \end{cases}$$

Remark 5.22. It is to be noted that $I_A(s)$ is also a random variable. Therefore, theoretically, we can view calculating probability as calculating the expectation of an indicator function:

$$E(I_A) = (1)P(I_A = 1) = P(A).$$

This is useful in proving theorems related to random variables.

Definition 5.23. (Variance) The variance of X, denoted by $Var(X)$, is defined by

$$Var(X) = E[(X - \mu)^2].$$

It is a measure of scattering (or spread) of values of X around its expected value, μ.

Definition 5.24. (Standard deviation) The standard deviation of X, denoted by σ_X or $SD(X)$, is defined as

$$\sigma_X = \sqrt{Var(X)}.$$

Proposition 5.25. *An alternative formula for variance is given by*

$$Var(X) = E(X^2) - [E(X)]^2.$$

Proof.

$$\begin{aligned} Var(X) = E[(X - \mu)^2] &= \sum_{x \in R(X)} (x - \mu)^2 p_X(x) \\ &= \sum_{x \in R(X)} x^2 p_X(x) - 2\mu \sum_{x \in R(X)} p_X(x) + \mu^2 \sum_{x \in R(X)} p_X(x) \\ &= E(X^2) - 2\mu(\mu) + \mu^2(1) = E(X^2) - \mu^2, \end{aligned}$$

where $\mu = E(X)$. □

Corollary 5.26. *For all discrete random variables X, we have $E(X^2) \geq [E(X)]^2$.*

Proof. Since $Var(X) = E[(X-\mu)^2]$, where $(X-\mu)^2$ is non-negative, we must have that $Var(X) \geq 0$. Therefore, by proposition 5.25, we have

$$E(X^2) - [E(X)]^2 \geq 0 \implies E(X^2) \geq [E(X)]^2.$$

We therefore conclude our proof. □

Remark 5.27. We have $Var(X) = 0$ if and only if X is a degenerate random variable. That is, a random variable taking only one value.

Example 5.28. I_S and $I_\emptyset$ are both degenerate random variables, where S is the sample space.

Definition 5.29. (Bernoulli random variable) Suppose X takes only two values 0 and 1 with $P(X=0) = 1-p$ and $P(X=1) = p$. We call X a Bernoulli random variable of parameter p, denoted by $X \sim Be(p)$.

Remark 5.30. The random variable defined in example 5.3 is a Bernoulli random variable of parameter $\frac{1}{2}$, i.e., $Be(\frac{1}{2})$.

Remark 5.31. Some textbooks use the term Bernoulli trials for such random variables.

Definition 5.32. (Binomial random variable) If we perform Bernoulli experiments independently under identical conditions n times and define

$$X = \text{the number of successes in } n \text{ Bernoulli trials of parameter } p,$$

we say that X is a binomial random variable and write $X \sim Bin(n,p)$.

Definition 5.33. (Geometric random variable) Define X to be the number of Bernoulli trials of parameter p required to obtain the first success. We say that X is a geometric random variable, written as $X \sim Geom(p)$.

Definition 5.34. (Negative binomial random variable) Define X to be the number of Bernoulli trials of parameter p required to obtain r successes. We say that X is a negative binomial random variable, denoted by $NB(r,p)$.

Proposition 5.35. *Let $X \sim Bin(n,p)$, $Y \sim Geom(p)$, $W \sim NB(r,p)$. We have*

1) $E(X) = np$, $E(Y) = \frac{1}{p}$, $E(W) = \frac{r}{p}$;

2) $Var(X) = np(1-p)$, $Var(Y) = \frac{1-p}{p^2}$, $Var(W) = \frac{r(1-p)}{p^2}$.

Proof. We only present the proof for 1) here since the proof for 2) will be more or less similar. By combinatoric analysis, we have

$$p_X(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} = \binom{n}{x}p^x(1-p)^{n-x}$$

$$p_Y(y) = (1-p)^{y-1}p \qquad .$$

$$p_W(w) = \binom{w-1}{r-1}p^r(1-p)^{w-r}$$

Therefore, we have

$$\begin{aligned}
E(X) &= \sum_{x=1}^{n} x p_X(x) \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!}p^x(1-p)^{n-x} \\
&= np\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!}p^{x-1}(1-p)^{n-x} \\
&= np[p+(1-p)]^{n-1} \\
&= np;
\end{aligned}$$

$$\begin{aligned}
E(Y) &= \sum_{y=1}^{\infty} yp(1-p)^{y-1} \\
&= p\frac{d}{dp}\left(-\sum_{y=1}^{\infty}(1-p)^y\right) \\
&= p\frac{d}{dp}\left(1-\frac{1}{p}\right) \\
&= \frac{1}{p}
\end{aligned}$$

To find $E(W)$, we will use the following lemma:

Lemma 5.36. *Let $\{X_i\}_{i\in I_n}$ be a collection of independent geometric random variables, where $I_n := \{r \in \mathbb{Z} : 1 \leq r \leq n\}$, and $Y := \sum_{i\in I_n} X_i$. We have $E(Y) = nE(X)$.*

By viewing $W = \sum_{i=1}^{r} Y_i$, where Y_i are independent geometric random variable of parameter p, we conclude that $E(W) = E(\sum_{i=1}^{r} Y_i) = \sum_{i=1}^{r} E(Y_i) = \frac{r}{p}$. □

Remark 5.37. Different materials may have different definitions for $Geom(p)$ and $NB(n, r)$. It is thus important to check what definitions the lecturer uses to define these two variables.

Definition 5.38. (Poisson distribution) A random variable is said to have a Poisson distribution with parameter λ if X takes non-negative values with probabilities given by

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

Proposition 5.39. *Let $X \sim Poisson(\lambda)$. We have $E(X) = Var(X) = \lambda$.*

Proof. Recall that

$$e^x = \sum_{r=0}^{\infty} \frac{x^r}{r!}.$$

By the definition of expectation, we have

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} k p_X(k) \\ &= \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

Similarly, we have $E(X^2) = \lambda(\lambda + 1)$, which gives us

$$Var(X) = E(X^2) - [E(X)]^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda.$$

□

Proposition 5.40. *Let $X \sim Bin(n,p)$. When n is large and p is small, we can take $\lambda = np$, and X can be approximated by $Poisson(\lambda)$.*

Proof. Take $\lambda = np$. We have

$$\begin{aligned} p_X(x) &= \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\ &= \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\prod_{r=n-x+1}^{n} r}{x!}\left(\frac{\lambda}{n}\right)^x\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} \\ &= \frac{\lambda^x}{x!}\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k}\prod_{r=n-x+1}^{n}\left(\frac{r}{n}\right) \end{aligned}$$

When n is very large and λ is moderate, we have

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \left(1-\frac{\lambda}{n}\right)^k \approx 1, \prod_{r=n-x+1}^{n}\left(\frac{r}{n}\right) \approx 1,$$

which then gives us

$$p_X(x) \approx \frac{e^{-\lambda}\lambda^x}{x!}.$$

□

Remark 5.41. The author feels that $\prod_{r=n-x+1}^{n}\left(\frac{r}{n}\right) \approx 1$ works well for smaller x but not the large ones, it may be possible to further impose the condition that x must also be small.

Remark 5.42. As a working rule, use the Poisson distribution if $p < 0.1$ and put $\lambda = np$. If $p > 0.9$, we then put $\lambda = n(1-p)$ and work in terms of "failure".

Remark 5.43. In the lecture notes, a lot of 'random' random variables are introduced. For the sake of the succinctness of this set of notes, the author intentionally excludes some of the random variables introduced in the lecture notes. Yet, all discrete random variables excluded can be derived by appropriate combinatoric analysis, while all continuous random variables excluded will only be used in more advanced analysis.

6. Continuous Random Variables

From this section onwards, proving certain results will need tremendous knowledge from other fields of mathematics, some of which are currently beyond the author's ability. Therefore, some proofs will be omitted, and those results may be taken as granted.

Definition 6.1. (Continuous random variable) We say that X is a continuous random variable if there exists a nonnegative function f_X, defined for all real $x \in \mathbb{R}$, such that

$$P(a < X \leq b) = \int_a^b f_X(x)\ dx,$$

for $-\infty < a < b < \infty$. The function f_X is called the probability density function (pdf) of the random variable X.

Remark 6.2. The correct interpretation of pdf should be

$$f_X(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}.$$

Remark 6.3. The above definition directly implies that $P(X = a) = 0$ for all $a \in \mathbb{R}$, given that X is a continuous random variable. However, this does not suggest that it is impossible for $X = a$ to occur.

Example 6.4. Let X be the number arbitrarily and uniformly chosen from the interval $[0, 1]$. With the above definition, we have $P(X = 1) = 0$ and $P(0 \leq X < 1) = 1$. Yet, it is still possible for X to be 1. Therefore, events with probability 0 may still occur, and events with probability 1 may still not occur.

Definition 6.5. (Cumulative distribution function/cdf) We define the (cumulative) distribution function of X by

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

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

Proposition 6.6. *Let X be a continuous random variable. Then, we have $f_X(x) = \frac{d}{dx} F_X(x)$.*

Proof. This is a direct result implied by the fundamental theorem of calculus:

$$\begin{aligned}\frac{d}{dx}F_X(x) &= \lim_{h\to 0}\frac{\int_{-\infty}^{x+h} f_X(t)\ dt - \int_{-\infty}^{x} f_X(t)\ dt}{h}\\ &= \lim_{h\to 0}\frac{\int_{x}^{x+h} f_X(t)\ dt}{h}\\ &= \lim_{h\to 0}\frac{G(x+h)-G(x)}{h}\\ &= G'(x)\\ &= f_X(x),\end{aligned}$$

where $G(x)$ is the antiderivative of $f_X(t)$. □

Definition 6.7. (Expectation) Let X be a continuous random variable with pdf f_X, then

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

Definition 6.8. (Variance) Let X be a continuous random variable with pdf f_X, then

$$Var(X) = \int_{-\infty}^{\infty} (x - E(x))^2 f_X(x)\ dx.$$

Remark 6.9. You may notice the similarity between the definitions of expectation and variance of discrete random variables and those of continuous random variables.

Proposition 6.10. *(Tail-sum formula) Suppose X is a nonnegative continuous random variable, then*

$$E(X) = \int_0^{\infty} P(X > x)\ dx = \int_0^{\infty} P(X \geq x)\ dx.$$

Proof. The second equality is trivial since X is a continuous random variable. We thus only present the proof for the first equality here. By the definition of expectation, we have

$$E(X) = \int_0^{\infty} x f_X(x)\ dx = \int_0^{\infty} x F_X'(x)\ dx.$$

Noticing that

$$x = \int_0^{x} 1\ dt,$$

we have

$$\begin{aligned} E(X) &= \int_0^\infty \left(\int_0^x 1 \, dt \right) F_X'(x) \, dx \\ &= \int_0^\infty \int_0^x F_X'(x) \, dt \, dx \\ &= \int_0^\infty \int_t^\infty F_X'(x) \, dx \, dt \\ &= \int_0^\infty 1 - F_X(t) \, dt \\ &= \int_0^\infty P(X \geq t) \, dt. \end{aligned}$$

We thus conclude our proof. □

Remark 6.11. The proof presented in the lecture notes makes use of the indicator, which unnecessarily complicates the proof in my opinion. There is an error in that proof also.

Proposition 6.12. *Similar to the case of discrete random variables, we have*

1) $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx$ *for any bounded function* $g(x)$*, i.e., there exists a* $M > 0$ *such that* $|g(x)| < M$ *for all* $x \in \mathbb{R}$*.;*

2) $E(aX + b) = aE(X) + b$*;*

3) $Var(X) = E(X^2) - [E(X)]^2$.

Proof. We will only present the proof for 1). Since $g(x)$ is bounded, there exists a $M > 0$ such that $|g(x)| < M$, which implies that $g(x) + M$ must be non-negative. Let $h(x) = g(x) + M$. By the tail-sum formula, we have

$$\begin{aligned} E[h(X)] &= \int_0^\infty P(h(X) \geq t) \, dt \\ &= \int_0^\infty \int_{x:h(x)\geq t} f_X(x) \, dx \, dt \\ &= \int_{x:h(x)\geq 0} \int_0^{h(x)} f_X(x) \, dt \, dx \\ &= \int_{x:h(x)\geq 0} h(x) f_X(x) \, dx. \end{aligned}$$

Since $h(x)$ is non-negative, we have $h(x) \geq 0$ for all $x \in \mathbb{R}$, so we have

$$\begin{aligned} E[h(X)] = \int_{x:h(x)\geq 0} h(x) f_X(x)\ dx &\iff E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\ dx \\ &\iff E[g(X)+M] = \int_{-\infty}^{\infty} [g(x)+M] f_X(x)\ dx \\ &\iff E[g(X)] + M = M + \int_{-\infty}^{\infty} g(x) f_X(x)\ dx \\ &\iff E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\ dx \end{aligned}$$

as per required. □

Remark 6.13. Technically, we need another lemma which says that $E(X+M) = E(X) + M$, where M is a constant. Yet, this should be obvious since $Y := X + M$ defines a bijection.

Definition 6.14. (Uniform distribution) A random variable X is said to be uniformly distributed over the interval (a, b) if its probability density function is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{otherwise} \end{cases}.$$

We denote this by $X \sim U(a, b)$.

Definition 6.15. (Normal distribution) A random variable X is said to be normally distributed with parameters μ and σ^2 if its probability density function is given by

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

for $-\infty < x < \infty$. We denote this by $X \sim N(\mu, \sigma^2)$.

Definition 6.16. (Standard normal random variable) A normal random variable is called a standard normal random variable when $\mu = 0$ and $\sigma = 1$. We denote the standard normal random variable by Z, i.e., $Z \sim N(0, 1)$. Its pdf is usually denoted by ϕ and its cdf by Φ.

Proposition 6.17. (Properties of the standard normal distribution)

1) $P(Z \geq 0) = P(Z \leq 0) = 0.5$ *since it is symmetric about the y-axis;*

2) $-Z \sim N(0,1)$ *since it is symmetric about the y-axis.*

Definition 6.18. (Exponential distribution) A random variable is said to be exponentially distributed with parameter $\lambda > 0$ if its pdf is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda t} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

It is denoted by $X \sim Exp(\lambda)$. In addition, we have $E(X) = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$.

Proposition 6.19. *(Memoryless property) Let T be the exponential distribution with parameter λ. For $s, t > 0$, we have*

$$P(T > s + t | T > s) = P(T > t).$$

Proof. The cdf of $X \sim Exp(\lambda)$ is given by

$$F_X(x) = \int_0^x \lambda e^{-\lambda t}\, dt = \left[-e^{-\lambda t}\right]_0^x = 1 - e^{-\lambda x}$$

By the definition of conditional probability, we have

$$\begin{aligned} P(T > s + t | T > s) &= \frac{P(T > s + t)}{P(T > s)} \\ &= \frac{1 - [1 - e^{-\lambda(s+t)}]}{1 - [1 - e^{-\lambda s}]} \\ &= e^{-\lambda t} = P(T > t). \end{aligned}$$

We therefore conclude our proof. □

Definition 6.20. (Gamma distribution) A random variable X is said to have a gamma distribution with parameters (α, λ), denoted by $Gamma(\alpha, \lambda)$, if its pdf is given by

$$f_X(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha - 1}}{\Gamma(\alpha)} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

where $\lambda, \alpha > 0$ and $\lambda(\alpha)$, known as the gamma function, is defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1}\, dy.$$

In addition, we have $E(X) = \frac{\alpha}{\lambda}$ and $Var(X) = \frac{\alpha}{\lambda^2}$.

Proposition 6.21. *We have following results regarding gamma distribution:*

1) $Gamma(1, \lambda) = Exp(\lambda)$*;*

2) if $X_i \sim Exp(\lambda)$*, then* $\sum_{r=1}^n X_r \sim Gamma(n, \lambda)$*;*

3) If $X \sim Gamma(\frac{n}{2}, \frac{1}{2})$*, then* $X \sim \chi^2(n)$*;*

4) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Remark 6.22. If identical events (whose occurrence follows Poisson distributions of parameter λ) are occurring randomly and independently in time, then the amount of time one has to wait until a total of n events has occurred is a random variable which follows a Gamma Distribution with parameters (n, λ). With the above propositions, it explains the significance of $Exp(\lambda)$.

Theorem 6.23. *(De Moivre-Laplace Limit Theorem) Suppose that* $X \sim Bin(n, p)$*. Then, for any* $a < b$*, we have*

$$P\left(a < \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) \to \Phi(b) - \Phi(a)$$

as $n \to \infty$*, where* $\Phi(z) = P(Z \leq z)$ *with* $Z \sim N(0, 1)$.

Remark 6.24. The above theorem implies that when the number of Bernoulli trials is large enough, $X \sim Bin(n, p)$ can be approximated by $N(np, np(1-p))$, i.e., the normal distribution with the same mean and variance.

Remark 6.25. Generally, when we have $np(1-p) \geq 10$, the normal approximation of the binomial random variable is quite satisfying.

Proposition 6.26. *(Continuity-correction) If $X \sim Bin(n,p)$, then we have*

$$P(X=k) = P\left(k-\frac{1}{2} < x < k+\frac{1}{2}\right),$$
$$P(X \geq k) = P\left(X \geq k-\frac{1}{2}\right),$$
$$P(X \leq k) = P\left(X \leq k+\frac{1}{2}\right).$$

Remark 6.27. Since binomial random variables are discrete while normal random variables are continuous, to adjust this difference, we need the continuity-correction above.

Definition 6.28. (Chi-square distribution) Let $X \sim N(0,1)$. We define $Y = X^2$ as a chi-square random variable of degree 1.

Theorem 6.29. *Let X be a continuous random variable having pdf f_X. If $g(x)$ is a strictly monotonic (increasing or decreasing), differentiable function of X, then the random variable Y defined by $Y = g(X)$ has a pdf given by*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left|\frac{d}{dy} g^{-1}(y)\right| & \text{if } y = g(x) \text{ for some } x; \\ 0 & \text{otherwise.} \end{cases},$$

where $g^{-1}(y)$ is defined to be equal to the value of X such that $g(x) = y$.

Proof. WLOG, we assume that $g(x)$ is an increasing function. Suppose $y = g(x)$ for some x. Then, with $Y = g(X)$, we have

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

Differentiating both sides with respect to y, we have

$$f_Y(y) = f_X(g^{-1}(y)) \left[\frac{d}{dy} g^{-1}(y)\right],$$

which agrees with the form given in the theorem. □

7. Jointly Distributed Random Variables

Definition 7.1. (Jointly discrete random variables) For any two random variables X and Y defined on the same sample space, we define the joint distribution function of X and Y, denoted by $F_{X,Y}(x, y)$, by

$$F_{X,Y}(x, y) := P(X \leq x, Y \leq y)$$

for all $x, y \in \mathbb{R}$. We say that X and Y are jointly discrete random variables if such $F_{X,Y}$ exists.

Remark 7.2. The definition does not require X and Y to be independent. Hence, we should understand $F_{X,Y}(x, y)$ as purely the probability of $X \leq x$ and $Y \leq y$ happening together instead of the product of $P(X \leq x)$ and $P(Y \leq y)$.

Definition 7.3. (Marginal distribution function of X) The distribution function of X can be obtained from $F_{X,Y}(x, y)$ by

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y).$$

We then call $F_X(x)$ the marginal distribution function of X.

Remark 7.4. As $y \to \infty$, $P(Y \leq y)$ will tend to 1, which suggests that eventually, $Y \leq y$ will definitely happen. Therefore, the probability of $X \leq x$ and $Y \leq y$ happening together is just the probability of $X \leq x$, which is $F_X(x)$.

Proposition 7.5. *Let $a, b, a_1 < a_2, b_1 < b_2$ be real numbers. We have*

1) $P(X > a, Y > b) = 1 - F_X(a) - F_Y(b) + F_{X,Y}(a, b)$;
2) $P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$.

Remark 7.6. You may want to plot the corresponding regions on a $\mathbb{R}^2$ plane to make sense of the above propositions.

Definition 7.7. (Joint probability mass function) Let X and Y be jointly discrete random variables. We define the joint probability mass function of X and Y as

$$p_{X,Y}(x, y) := P(X = x, Y = y).$$

We can define the marginal pmf in the same way as the definition 7.3.

Proposition 7.8. *Let X, Y be jointly discrete random variables. We have*

1) $P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = \sum_{a_1<x\leq a_2} \sum_{b_1<y\leq b_2} p_{X,Y}(x,y)$;

2) $F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = \sum_{x\leq a} \sum_{y\leq b} p_{X,Y}(x,y)$;

3) $P(X > a, Y > b) = \sum_{x>a} \sum_{y>b} p_{X,Y}(x,y)$.

Definition 7.9. (Jointly continuous random variables) We say that X and Y are jointly continuous random variables if there exists a function (known as the joint probability density function of X and Y) defined for all $x, y \in \mathbb{R}$, such that for every $C \subseteq \mathbb{R}^2$, we have

$$P((X,Y) \in C) := \iint_{(x,y)\in C} f_{X,Y}(x,y)\ dx\ dy.$$

Proposition 7.10.

1) *Let $A, B \subseteq \mathbb{R}$, and $C = A \times B$. We have*

$$P((X,Y) \in C) = P(X \in A, Y \in B) = \int_A \int_B f_{X,Y}(x,y)\ dy\ dx.$$

2) *Let $a, b \in \mathbb{R}$. We have*

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

3) *Let $a_1, a_2, b_1, b_2 \in \mathbb{R}$. We have*

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)\ dy\ dx.$$

Remark 7.11. As a result of this,

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

Definition 7.12. (Marginal probability density function) The marginal pdf of X is given by

$$f_X(x) := \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy.$$

Definition 7.13. (Independent events) Two random variables X and Y are said to be independent if

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

for any $A, B \subseteq \mathbb{R}$. Random variables that are not independent are said to be dependent.

Proposition 7.14. *The following three statements are equivalent:*

1) Random variables X and Y are independent;

2) For all $x, y \in \mathbb{R}$, we have $f_{X,Y}(x, y) = f_X(x)f_Y(y)$;

3) For all $x, y \in \mathbb{R}$, we have $F_{X,Y}(x, y) = F_X(x)F_Y(y)$.

Proof. 1) $\iff$ 2) by the definition of independent events.

2) $\implies$ 3) by direct integrating both sides of $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ with respect to x, y.

3) $\implies$ 2) Suppose $X \times Y \in [a, b] \times [c, d]$. From 3), we have

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \implies \int_a^x \int_c^y f_{X,Y}(s, t)\, dt\, ds = \int_a^x f_X(s)\, ds \int_c^y f_Y(t)\, dt.$$

Differentiating both sides with respect to y, we have

$$\int_a^x f_{X,Y}(s, y)\, ds = f_Y(y) \int_a^x f_X(s)\, ds.$$

Differentiating both sides with respect to x, we have

$$f_{X,Y}(x, y)\, ds = f_Y(y)f_X(x),$$

which gives us the desired result. □

Proposition 7.15. *Random variables X and Y are independent if and only if there exist functions $g, h : \mathbb{R} \mapsto \mathbb{R}$ such that for all $x, y \in \mathbb{R}$, we have*

$$f_{X,Y}(x, y) = h(x)g(y).$$

Proof. We only present the proof for the continuous case.

($\implies$) Suppose X and Y are independent. We have

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),$$

so we can just take $h(x) = f_X(x)$ and $g(y) = f_Y(y)$.

($\impliedby$) Suppose we have $f_{X,Y}(x,y) = h(x)g(y)$. We aim to prove that X and Y are independent. Let $C_1 := \int_{-\infty}^{\infty} h(x)\ dx$ and $C_2 := \int_{-\infty}^{\infty} g(y)\ dy$. From our hypothesis, we have

$$\begin{aligned} 1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x)g(y)\ dx\ dy \\ &= \int_{-\infty}^{\infty} h(x)\ dx \int_{-\infty}^{\infty} g(y)\ dy \\ &= C_1C_2. \end{aligned}$$

Furthermore, by the definition of marginal pdf, we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy = \int_{-\infty}^{\infty} h(x)g(y)\ dy = h(x)C_1, \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx = \int_{-\infty}^{\infty} h(x)g(y)\ dx = g(y)C_2. \end{aligned}$$

Therfore, we have

$$f_X(x)f_Y(y) = C_1C_2h(x)g(y) = h(x)g(y) = f_{X,Y}(x,y),$$

which suggests that X, Y are independent. □

Proposition 7.16. (*Convolution*) *Assume that X, Y are independent jointly continuous random variables. The cdf F_{X+Y} is called the convolution of the distributions F_X and F_Y:*

$$\begin{aligned} F_{X+Y}(a) &= \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)\ dy = \int_{-\infty}^{\infty} F_Y(a-x)f_X(x)\ dx, \\ f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)\ dy = \int_{-\infty}^{\infty} f_Y(a-x)f_X(x)\ dx. \end{aligned}$$

Proof. We have

$$\begin{aligned} F_{X+Y}(a) = P(X+Y \leq a) &= \iint_{x+y\leq a} f_{X,Y}(x,y)\ dx\ dy \\ &= \iint_{x+y\leq a} f_X(x) f_Y(y)\ dx\ dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{a-y} f_X(x)\ dx \right] f_Y(y)\ dy \\ &= \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\ dy \end{aligned}$$

Differentiating both sides with respect to a gives us

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\ dy = \int_{-\infty}^{\infty} f_Y(a-x) f_X(x)\ dx.$$

We therefore conclude our proof. □

Proposition 7.17. *By the convolution formula, we have the following results,*

1) Let $X \sim Gamma(\alpha, \lambda)$ and $Y \sim Gamma(\beta, \lambda)$. We have

$$X + Y \sim Gamma(\alpha + \beta, \lambda).$$

2) Let X_i be independent normal random variables. We have

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right).$$

3) Let X and Y be Poisson random variables with parameter λ and μ respectively. We have

$$X + Y \sim Poisson(\lambda + \mu).$$

4) Let $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$. We have

$$X + Y \sim Bin(m + n, p).$$

Definition 7.18. (Conditional probability mass function) Let X, Y be jointly discrete random variables. The conditional probability of X given that $Y = y$ is defined by

$$p_{X|Y}(x|y) := P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

for all Y such that $p_Y(y) > 0$.

Definition 7.19. (Conditional distribution function) Let X, Y be jointly discrete random variables. The conditional distribution function of X given that $Y = y$ is defined by

$$F_{X|Y}(x|y) := P(X \leq x|Y = y) = \sum_{a \leq x} \frac{p_{X,Y}(x, y)}{p_Y(y)} = \sum_{a \leq x} p_{X|Y}(a|y).$$

Proposition 7.20. *If X is independent of Y, then $p_{X|Y}(x|y)$ is the same as $p_X(x)$ for Y satisfying $p_Y(y) > 0$.*

Proof. Since X, Y are independent, we have $p_{X,Y}(x, y) = p_X(x)p_Y(y)$, which gives us

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x).$$

The assumption that $p_Y(y) > 0$ is necessary since otherwise, it is impossible to define $p_{X|Y}(x|y)$ at the first place. □

Definition 7.21. (Conditional probability density function) Let X, Y be jointly continuous random variables. The conditional probability of X given that $Y = y$ is defined by

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

for all Y such that $f_Y(y) > 0$.

Definition 7.22. (Conditional distribution function) Let X, Y be jointly continuous random variables. The conditional distribution function of X given that $Y = y$ is defined by

$$F_{X|Y}(x|y) := P(X \leq x|Y = y) = \int_{-\infty}^{x} f_{X|Y}(t|y)\, dt.$$

Proposition 7.23. *Assume that*

1) Let X and Y be jointly continuously distributed random variables with known joint pdf.

2) Let U and V be given functions of X and Y in the form:

$$U = g(X, Y), \quad V = h(X, Y),$$

and we can uniquely solve X and Y in terms of U and V, say $x = a(u, v)$ and $y = b(u, v)$.

3) The functions g and h have continuous partial derivatives at all points (x, y) and

$$J(x,y) := \begin{vmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{vmatrix} = \frac{\partial g}{\partial x}\frac{\partial h}{\partial y} - \frac{\partial g}{\partial y}\frac{\partial h}{\partial x} \neq 0$$

at all points (x,y).

Then, the joint probability density function of U and V is given by

$$f_{U,V}(u, v) = f_{X,Y}(x, y)|J(x, y)|^{-1},$$

where $x = a(u, v)$ and $y = b(u, v)$.

Proof. You may consider the above process as a change of coordinates. The detailed proof will be covered in MA2104. □

Proposition 7.24. *For jointly continuous random variables, the following three statements are equivalent:*

1) Random variables X, Y, and Z are independent.

2) For all $x, y, z \in \mathbb{R}$, we have $f_{X,Y,Z}(x, y, z) = f_X(x)f_Y(y)f_Z(z)$.

3) For all $x, y, z \in \mathbb{R}$, we have $F_{X,Y,Z}(x, y, z) = F_X(x)F_Y(y)F_Z(z)$.

Proposition 7.25. *Random variables X, Y, and Z are independent if and only if there exist functions $g_1, g_2, g_3 : \mathbb{R} \mapsto \mathbb{R}$ such that for all $x, y, z \in \mathbb{R}$, we have*

$$f_{X,Y,Z}(x, y, z) = g_1(x)g_2(y)g_3(z).$$

Proof. The proof is very similar to the one for the case of two random variables. □

8. Properties of Expectation

Proposition 8.1. *If $a \leq X \leq b$, then $a \leq E(X) \leq b$.*

Proof. We only present the proof for the discrete case. We have

$$\begin{aligned} a \leq X \leq b &\iff (\forall x \in R(X))(a \leq x \leq b) \\ &\iff \sum_{x \in R(X)} a p_X(x) \leq \sum_{x \in R(X)} x p_X(x) \leq \sum_{x \in R(X)} b p_X(x) \\ &\iff a \sum_{x \in R(X)} p_X(x) \leq E(X) \leq b \sum_{x \in R(X)} p_X(x) \\ &\iff a \leq E(X) \leq b, \end{aligned}$$

given that by definition, $\sum_{x \in R(X)} p_X(x) = 1$. □

Proposition 8.2. *Let X_1, Y_1 be jointly discrete, and X_2, Y_2 be jointly continuous. We have*

1) $E[g(X_1, Y_1)] = \sum_y \sum_x g(x, y) p_{X_1, Y_1}(x, y)$.
2) $E[g(X_2, Y_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\ dx\ dy$.

Corollary 8.3. *There are some important results directly implied by the above propositions.*

1) *If $g(x, y) \geq 0$ whenever $p_{X,Y}(x, y) > 0$, then $E[g(X, Y)] \geq 0$.*
2) $E[g(X, Y) + h(X, Y)] = E[g(x, y)] + E[h(x, y)]$.
3) *If jointly distributed random variables X and Y satisfy $X \leq Y$, then $E(X) \leq E(Y)$.*

Remark 8.4. Therefore, we can confidently say that $E(X + Y) = E(X) + E(Y)$.

Definition 8.5. (Random sample) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having distribution function F and expected value μ. Such a sequence of random variables is said to constitute a random sample from the distribution F.

Definition 8.6. (Sample mean) We define the sample mean $\overline{X}$ by

$$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_k,$$

where $\{X_k\}$ is a random sample.

Proposition 8.7. *(Boole's Inequality) Let $A_1, \ldots, A_n$ be events. We have*

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

Proof. Let I_i be the indicator of the event A_i for $1 \leq i \leq n$. We define

$$X := \sum_{k=1}^{n} I_k,$$
$$Y := I\left(\bigcup_{k=1}^{n} A_k\right).$$

By the definition of indicator, clearly, we have $Y \leq X$ for all x in the sample space. This is because

(1) if we have $x \in (\bigcup_{k=1}^{n} A_k)$, then we must have $x \in A_i$ for some i, which suggests that $Y = 1$ and $X \geq 1$;

(2) if we have $x \notin (\bigcup_{k=1}^{n} A_k)$, then we have $X = Y = 0$.

Therefore, by the property of expectation, we have

$$\begin{aligned} Y \leq X &\iff E(Y) \leq E(X) \\ &\iff P\left(I\left(\bigcup_{k=1}^{n} A_k\right) = 1\right) \leq E\left(\sum_{k=1}^{n} I_k\right) \\ &\iff P\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} E\left(I_k\right) \\ &\iff P\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} P(A_k), \end{aligned}$$

which is the desired inequality. □

Definition 8.8. (Covariance) The covariance of jointly distributed random variables X and Y, denoted by $Cov(X, Y)$, is defined by

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)],$$

where μ_X and μ_Y denote the means of X and Y respectively.

Remark 8.9. If $Cov(X,Y) \neq 0$, we say that X and Y are correlated. Otherwise, we say that X and Y are uncorrelated. It is to be noted that correlation does not imply causation. This means, even though two things are correlated, we cannot conclude that one is the cause of the other.

Proposition 8.10. *$Cov(X,Y) = E(XY) - E(X)E(Y)$.*

Proof. Let μ_X and μ_Y be the mean of X and Y respectively, both of which are unknown constants. We have

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

□

Proposition 8.11. *If X and Y are independent, then for any functions $g, h : \mathbb{R} \mapsto \mathbb{R}$, we have*

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

Proof. We only prove the discrete case here and leave the continuous case as an exercise.

$$\begin{aligned} E[g(X)h(Y)] &= \sum_x \sum_y g(x)h(y)p_{X,Y}(x,y) \\ &= \sum_x \sum_y g(x)h(y)p_X(x)p_Y(y) \\ &= \sum_x g(x)p_X(x) \sum_y h(y)p_Y(y) \\ &= E[g(X)]E[h(Y)]. \end{aligned}$$

□

Corollary 8.12. *If X and Y are independent, then $Cov(X,Y) = 0$.*

Remark 8.13. It is to be noted that $Cov(X, Y) = 0$ does not imply that X and Y are independent.

Example 8.14. Let X be a random variable taking either -1 or 1 with equal possibility, i.e., $P(X = -1) = P(X = 1) = 0.5$. We define Y such that if $X = -1$, then $Y = 0$, and if $X = 1$, then Y takes either -1 or 1 with probability 0.5 respectively. Clearly, X and Y are dependent. However, we have

$$\begin{aligned} E(X) &= \frac{1}{2}(-1) + \frac{1}{2}(1) = 0; \\ E(Y) &= \frac{1}{2}(0) + \frac{1}{4}(1) + \frac{1}{4}(-1) = 0; \\ E(XY) &= \frac{1}{2}(-1)(0) + \frac{1}{4}(1)(-1) + \frac{1}{4}(1)(1) = 0, \end{aligned}$$

which gives $Var(X, Y) = E(XY) - E(X)E(Y) = 0$.

Remark 8.15. More generally, if we have $P(Y = a|X = x) = P(Y = -a|X = x)$ for all $a \in \mathbb{R}, x \in \mathbb{R}$, we will have zero covariance. Yet, as long as we do not have $P(Y|X) = P(Y)$, X and Y are not independent.

Proposition 8.16. *There are some important properties of covariances:*

1) $Var(X) = Cov(X, X)$.

2) $Cov(X, Y) = Cov(Y, X)$.

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

Proposition 8.17. $Var(\sum_{k=1}^{n} X_k) = \sum_{k=1}^{n} Var(X_k) + 2 \sum_{1 \le i < j \le n} Cov(X_i, X_j)$.

Proof. This is a direct implication of proposition 8.16 3). □

Corollary 8.18. *Let X_1, ..., X_n be independent random variables. We have*

$$Var\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} Var(X_k).$$

Proof. This is a direct implication of proposition 8.17. □

Definition 8.19. (Correlation coefficient) The random coefficient of random variables X and Y, denoted by $\rho(X,Y)$, is defined by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Proposition 8.20. *We have* $-1 \leq \rho(X,Y) \leq 1$.

Proof. Let X and Y be random variables with variances σ_X^2 and σ_Y^2 respectively.
1) Consider $Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right)$:

$$\begin{aligned}
0 &\leq Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\
&= Var\left(\frac{X}{\sigma_X}\right) + Var\left(\frac{Y}{\sigma_Y}\right) + 2Cov\left[\left(\frac{X}{\sigma_X}\right), \left(\frac{Y}{\sigma_Y}\right)\right] \\
&= Var\left(\frac{X}{\sigma_X}\right) + Var\left(\frac{Y}{\sigma_Y}\right) + 2\left(E\left[\left(\frac{X}{\sigma_X}\right)\left(\frac{Y}{\sigma_Y}\right)\right] + E\left(\frac{X}{\sigma_X}\right) + E\left(\frac{Y}{\sigma_Y}\right)\right) \\
&= Var\left(\frac{X}{\sigma_X}\right) + Var\left(\frac{Y}{\sigma_Y}\right) + 2\frac{Cov(X,Y)}{\sigma_X\sigma_Y} \\
&= 1 + 1 + 2\rho(X,Y),
\end{aligned}$$

which implies that $\rho(X,Y) \geq -1$.
2) Similarly, by considering $Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)$, we conclude that $\rho(X,Y) \leq 1$.
Combining both inequalities gives us the desired result. □

Remark 8.21. If X and Y are independent, then $\rho(X,Y) = 0$. Yet, again, the converse does not hold, i.e., $\rho(X,Y) = 0$ does not imply that X and Y are independent.

Definition 8.22. (Conditional expectation) If X_1 and Y_1 are jointly distributed discrete random variables with $p_Y(y) > 0$, and X_2 and Y_2 are jointly distributed continuous random variables with $f_Y(y) > 0$, then

$$E[X_1|Y_1 = y] = \sum_x x p_{X|Y}(x|y)$$

$$E[X_2|Y_2 = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x,y)\, dx.$$

Remark 8.23. We sometimes use $E(X|Y)$ to denote a function of y whose value at $Y = y$ is given by $E(X|Y = y)$.

Proposition 8.24. *Similarly, we have*

$$E[g(X)|Y = y] = \begin{cases} \sum_x g(x)p_{X|Y}(x, y) & \text{if } X, Y \text{ are discrete;} \\ \int_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)\ dx & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases}$$

Corollary 8.25. *We have*

$$E\left[\sum_{k=1}^{n} X_k|Y = y\right] = \sum_{k=1}^{n} E[X_k|Y = y].$$

Proof. This is a direct implication of proposition 8.24. □

Proposition 8.26. *We have*

$$E[X] = E[E[X|Y]].$$

Proof. We will only prove the continuous case.

$$\begin{aligned} E[E(X|Y)] &= \int_{-\infty}^{\infty} E(X|Y = y)f_Y(y)\ dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} xf_{X|Y}(x, y)\ dx\right) f_Y(y)\ dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf_{X|Y}(x, y)f_Y(y)\ dx\ dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf_{X,Y}(x, y)\ dx\ dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf_{X,Y}(x, y)\ dy\ dx \\ &= \int_{-\infty}^{\infty} xf_X(x)\ dx \\ &= E(X). \end{aligned}$$

□

Lemma 8.27. $Var(X|Y) = E(X^2|Y) - [E(X|Y)]^2$.

Proposition 8.28. *(Law of Total Variance) We have*

$$Var(X) = E[Var(X|Y)] + Var[E(X|Y)].$$

Proof. By the definition of variance, we have

$$\begin{aligned} Var(X) &= E(X^2) - [E(X)]^2 \\ &= E(E(X^2|Y)) - [E(X)]^2 \\ &= E(Var(X|Y) + [E(X|Y)]^2) - [E(X)]^2 \\ &= E[Var(X|Y)] + E\left([E(X|Y)]^2\right) - [E(E(X|Y))]^2 \\ &= E[Var(X|Y)] + Var[E(X|Y)], \end{aligned}$$

which is the desired form. □

Definition 8.29. (Moment generating function) The moment generating function of random variable X, denoted by M_X, is defined as

$$M_X(t) = E[e^{tx}] = \begin{cases} \sum_x e^{tx} p_X(x) & \text{if } X \text{ is discrete with pmf } p_X(x); \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx & \text{if } X \text{ is continuous with pdf } f_X(x). \end{cases}$$

Proposition 8.30. *Let $M_X^{(n)}(t)$ be the n-th derivative of $M_X(t)$. We have*

$$E(X^n) = M_X^{(n)}(0).$$

Proof. This can be proven by simply finding the n-th derivative of $M_X(t)$. □

Proposition 8.31. *If X and Y are independent, then*

$$M_{X+Y}(t) = M_X(t) M_Y(t).$$

Proposition 8.32. *Let X and Y be random variables with their moment generating functions $M_X(t)$ and $M_Y(t)$ respectively. If there exists an $h > 0$ such that*

$$M_X(t) = M_Y(t)$$

for all $t \in (-h, h)$, then X and Y have the same distribution.

Definition 8.33. (Joint Moment Generating Functions) For any n random variables X_1, X_2, ..., X_n, the joint moment generating function, $M_{X_1,\ldots,X_n}(t_1,\ldots,t_n)$, is defined for all $t_1,\ldots,t_n \in \mathbb{R}$ by

$$M_{X_1,\ldots,X_n}(t_1,\ldots,t_n) = E\left[e^{\sum_{i=1}^{n} t_i X_i}\right].$$

Proposition 8.34. *If $X_1,\ldots,X_n$ are independent and identically distributed normal random variables with mean μ and variance σ^2, then the sample mean $\overline{X}$ and the sample variance S^2 are independent. We also have*

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

9. Limit Theorems

Proposition 9.1. *(Markov's inequality) Let X be a nonnegative random variable. For $a > 0$, we have*

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

Proof. For $a > 0$, we define the indicator I as

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

Since $X \geq 0$, clearly, we have

$$I \leq \frac{X}{a}.$$

Taking the expectation both sides, we have

$$P(X \geq a) = E(I) \leq E\left(\frac{X}{a}\right) = \frac{E(X)}{a}$$

as the desired result. □

Proposition 9.2. *(Chebyshev's inequality) Let X be a random variable with mean μ, then for $a > 0$, we have*

$$P(|X - \mu| \geq a) \leq \frac{Var(X)}{a^2}.$$

Proof. By applying the Markov's inequality to $|X - \mu|$ and a, we have

$$P(|X - \mu| \geq a) = P(|X - \mu|^2 \geq a^2) \leq \frac{E(|X - \mu|^2}{a^2} = \frac{Var(X)}{a^2}$$

as the desired inequality. □

Corollary 9.3. *If $Var(X) = 0$, then the random variable X is a constant.*

Remark 9.4. Since Chebyshev's inequality works for all random variables, it can be expected that $\frac{Var(X)}{a^2}$ may not be a very good upper bound.

Theorem 9.5. (The Weak Law of Large Numbers) *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, with common mean μ. Then, for any $\varepsilon > 0$, we have*

$$P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

Theorem 9.6. (The Central Limit Theorem) *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then, the distribution of*

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}}\, dt.$$

Definition 9.7. Let $Z_1, \ldots$ be a sequence of random variables having distribution function F_{Z_n} and moment generating function M_{Z_n}, for $n \geq 1$. Let Z be a random variable having distribution function F_Z and moment generating function M_Z. If $M_{Z_n} \to M_Z(t)$ for all t, then

$$F_{Z_n}(x) \to F_Z(x)$$

for all x at which $F_Z(x)$ is continuous.

Proposition 9.8. *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each having mean μ and variance σ^2. Then, for large n, the distribution of*

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$$

is approximately standard normal.

Theorem 9.9. *(The Strong Law of Large Numbers) Let X_1, X_2, ... be a sequence of independent and identically distributed random variables, each having a finite mean $\mu = E[X_i]$. We have with probability 1,*

$$\frac{\sum_{i=1}^n X_i}{n} \to \mu \text{ as } n \to \infty.$$

Proposition 9.10. *(One-sided Checbyshev's inequality) If X is a random variable with mean 0 and finite variance σ^2, then, for any $a > 0$, we have*

$$P(X \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

Definition 9.11. (Convex function) A function $g(x)$ is convex if it meets one of the following conditions:

1) for all $0 \leq p \leq 1$ and all $x_1, x_2 \in R_X$, we have

$$g(px_1 + (1-p)x_2) \leq pg(x_1) + (1-p)g(x_2).$$

2) $g(x)$ is a convex differentiable function of one variable iff

$$g(x) \geq g(y) + g'(y)(x - y)$$

for all x and y in the interval.

3) A twice differentiable function of one variable is convex over an interval iff its second derivative is non-negative.

Proposition 9.12. *(Jensen's Inequality) If $g(x)$ is a convex function, then*

$$E[g(X)] \geq g(E[X])$$

provided that the expectations exist and are finite.