Probability Theory (MA 590)

Class Notes

January – May, 2023

Instructor
Ayon Ganguly
Department of Mathematics
IIT Guwahati

Contents

Chapter 1

Probability

1.1 Probability

1.1.1 Classical Probability

As we know that the probability of an event A, denoted by $P(A)$, is defined by

$$P(A) = \frac{\text{Favourable number of cases to } A}{\text{Total number of cases}} = \frac{\#A}{\#S},$$

where S is the set of all possible outcomes. This definition is known as classical definition of probability. Note that this definition is only meaningful if number of elements in S is finite. As $A \subseteq S$, A is finite if S is finite. Let us consider the following examples.

Example 1.1. Let a fair die is rolled. What is the probability of getting three on upper face? It is easy to see that the required probability is 1/6. ||

Example 1.2. Consider a target comprising of three concentric circles of radii 1/3, 1, and $\sqrt{3}$ feet. Consider the event that a shooter hits inside the inner circle and its' probability. Let A be an event that the shooter hits inside the inner circle. Then

$$S = \mathbb{R}^2 \quad \text{and} \quad A = \left\{ x \in \mathbb{R}^2 : |x| \leq \frac{1}{3} \right\}.$$

In this case both A and S are infinite and therefore probability of A can not be found using classical definition of probability. ||

Here we will try to provided a general definition of probability such that we can apply the new definition for larger class of problems, like Example 1.2. Note that probability can be viewed as a function where the argument of the function is a set and output is a real number. To give the new definition of probability, we will use three basic properties (will be discussed) of the classical definition of probability, and we will say that a function which satisfy these three properties is called a probability or a probability function. Of course, we need to define the domain of the function properly.

Definition 1.1 (Set Function)**.** *A function which takes a set as its' argument is called a set function.*

1.1.2 Countable and Uncountable Sets

For further discussion, we need the concepts of countable and uncountable sets. The definitions and some properties of countable and uncountable sets are given in this subsection. You must have read these concepts in analysis course and therefore it is a recapitulation.

Definition 1.2. *We say that two sets A and B are equivalent if there exists a bijection from A to B. We denote it by $A \sim B$.*

Definition 1.3. *For any positive integer n, let $J_n = \{1, 2, \ldots, n\}$ and $\mathbb{N}$ be the set of all positive integers (natural numbers). For any set A, we say:*

(a) A is finite if $A = \phi$ or $A \sim J_n$ for some $n \in \mathbb{N}$. n is said to be the cardinality of A or number of elements in A.

(b) A is infinite if A is not finite.

(c) A is countable if $A \sim \mathbb{N}$.

(d) A is atmost countable if A is finite or countable.

(e) A is uncountable if A is neither finite nor countable.

Example 1.3. The set of all integers, $\mathbb{Z}$, is countable. Consider the function $f : \mathbb{N} \to \mathbb{Z}$ given by

$$f(n) = \begin{cases} \frac{n}{2} & \text{if } n \text{ even} \\ -\frac{n-1}{2} & \text{if } n \text{ odd.} \end{cases}$$

It is easy to see that $f(\cdot)$ is a bijection from $\mathbb{N}$ to $\mathbb{Z}$. Therefore, $\mathbb{Z}$ is countable. ||

Remark 1.1. A finite set cannot be equivalent to any of its proper subset. However, this is possible for an infinite set. For example, consider the bijection $f : \mathbb{N} \to 2\mathbb{N}$ defined by

$$f(n) = 2n.$$

Here, $2\mathbb{N}$ is a proper subset of $\mathbb{N}$ and $f(\cdot)$ is a bijection from $\mathbb{N}$ to $2\mathbb{N}$. Therefore, $\mathbb{N}$ and $2\mathbb{N}$ are equivalent. †

Remark 1.2. If a set is countable, then it can be written as a sequence $\{x_n\}_{n\geq 1}$ of distinct terms. †

Theorem 1.1. *Every infinite subset of a countable set A is countable.*

Theorem 1.2. *Let $\{E_n\}_{n\geq 1}$ be a sequence of atmost countable sets and $S = \bigcup_{n=1}^{\infty} E_n$. Then S is atmost countable.*

Theorem 1.3. *Let $A_1, A_2, \ldots, A_n$ be atmost countable sets. Then $B_n = A_1 \times A_2 \times \ldots \times A_n$ is atmost countable.*

Corollary 1.1. *The set of rationals, $\mathbb{Q}$, is countable.*

Theorem 1.4. *The set, A, of all binary sequences is uncountable.*

Corollary 1.2. $[0, 1]$ *is uncountable.*

Corollary 1.3. $\mathbb{R}$ *is uncountable.*

Corollary 1.4. $\mathbb{Q}^c$ *is uncountable.*

Corollary 1.5. *Any interval is uncountable.*

1.1.3 Axiomatic Probability

To define the probability under more general framework, we need the concepts of random experiment, sample space, σ-field. These concepts are needed to define the domain of the probability function adequately.

Definition 1.4 (Random Experiment)**.** *An experiment is called a random experiment if it satisfies the following three properties:*

1. *All the outcomes of the experiment is known in advance.*
2. *The outcome of a particular performance of the experiment is not known in advance.*
3. *The experiment can be repeated under identical conditions.*

Note that according to the definition of a random experiment, we know all possible outcomes before hand, and hence we can make a list of all possible outcomes. This list is called sample space. The third condition in the definition of random sample is some what hypothetical in the sense that we will in general assume that the third condition is satisfied (if not very absurd to assume).

Definition 1.5 (Sample Space)**.** *The collection of all possible outcomes of a random experiment is called the sample space of the random experiment. It will be denoted by $\mathcal{S}$.*

Example 1.4. A toss of a coin is a random experiment as all the conditions of the definition of random experiment hold true. In this case, the sample space is $\mathcal{S} = \{H,\, T\}$. The sample space is finite in this example. ||

Example 1.5. Tossing a coin until the first head appears is also a random experiment with sample space $\mathcal{S} = \{H,\, TH,\, TTH,\, \ldots\}$. In this case, the sample space is countably infinite. ||

Example 1.6. The experiment of measuring the height of a student is a random experiment with sample space $\mathcal{S} = (0,\, \infty)$. Here the sample space is uncountable. ||

Definition 1.6 (σ-algebra or σ-field)**.** *A collection, $\mathcal{F}$, of subsets of $\mathcal{S}$ is called a σ-algebra or a σ-field if it satisfy the following properties:*

1. $\mathcal{S} \in \mathcal{F}$.
2. $A \in \mathcal{F}$ *implies* $A^c \in \mathcal{F}$.
3. $A_1,\, A_2,\, \ldots \in \mathcal{F}$ *implies* $\displaystyle\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The first condition in the definition of σ-field implies that a σ-field is always non-empty. The second condition is called closed under complementation. Note that here $A^c = \mathcal{S} - A$, *i.e.*, the complementation is with respect to the sample space. The third condition of the definition of σ-field is called closed under countable union. Thus, by definition, a σ-field is closed under complementation and countable union.

Definition 1.7 (Event)**.** *A set E is said to be an event with respect to a σ-field $\mathcal{F}$ if $E \in \mathcal{F}$. We will say "the event E occurs" if the outcomes of a performance of the random experiment is in E.*

Example 1.7 (Continuation of Example 1.4)**.** Consider following three classes of subsets of $\mathcal{S}$. $\mathcal{F}_1 = \{\emptyset,\, \mathcal{S},\, \{H\},\, \{T\}\}$, $\mathcal{F}_2 = \{\emptyset,\, \mathcal{S}\}$, and $\mathcal{F}_3 = \{\emptyset,\, \mathcal{S},\, \{H\}\}$. Here we will show that $\mathcal{F}_1$ and $\mathcal{F}_2$ are σ fields, but $\mathcal{F}_3$ is not a σ-field.

Note that $\mathcal{S} \in \mathcal{F}_1$ and for any $A \in \mathcal{F}_1$, $A^c \in \mathcal{F}_1$. Hence, it is easy to see that the first two conditions of the definition of σ-field are hold true. For the third condition, let $A_1,\, A_2,\, \ldots \in \mathcal{F}_1$.
Case I: If $A_i = \mathcal{S}$ for at least one $i \in \mathbb{N}$, $\bigcup_{i=1}^{\infty} A_i = \mathcal{S} \in \mathcal{F}_1$.
Case II: If $A_i = \emptyset$ for all $i \in \mathbb{N}$, $\bigcup_{i=1}^{\infty} A_i = \emptyset \in \mathcal{F}_1$.
Case III: If $A_i = \{H\}$ for at least one $i \in \mathbb{N}$ and rest of $A_i = \emptyset$, $\bigcup_{i=1}^{\infty} A_i = \{H\} \in \mathcal{F}_1$.
Case IV: If $A_i = \{T\}$ for at least one $i \in \mathbb{N}$ and rest of $A_i = \emptyset$, $\bigcup_{i=1}^{\infty} A_i = \{T\} \in \mathcal{F}_1$.
Case V: If $A_i = \{H\}$ for at least one $i \in \mathbb{N}$ and $A_j = \{T\}$ for at least one $j \in \mathbb{N}$, $\bigcup_{i=1}^{\infty} A_i = \mathcal{S} \in \mathcal{F}_1$.
These are the exhaustive cases and in all the cases, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}_1$. Hence, $\mathcal{F}_1$ is a σ-field on the subsets of $\mathcal{S}$. Note that $\mathcal{F}_1$ is power set of $\mathcal{S}$.

It is easy to see that $\mathcal{F}_2$ is σ-field and therefore left as a practice problem. To show that $\mathcal{F}_3$ is not a σ-field, we need to show that at least one of the three conditions is not true. It is very easy to check that the second condition is not true as $\{H\} \in \mathcal{F}_3$, but $\{H\}^c = \{T\} \notin \mathcal{F}_3$. ||

Example 1.8 (Continuation of Example 1.5)**.** Consider $\mathcal{F} = \mathcal{P}(\mathcal{S})$, the power set of $\mathcal{S}$. Clearly, $\mathcal{S} \in \mathcal{F}$. For any $A \in \mathcal{F}$, A^c is a subset of $\mathcal{S}$ and hence belongs to $\mathcal{F}$. For any countable collection of sets $A_1,\, A_2,\, \ldots \in \mathcal{F}$, $\bigcup_{i=1}^{\infty} A_i$ is a subset of $\mathcal{S}$ and belongs to $\mathcal{F}$. Hence, $\mathcal{F}$ is a σ-field on the subsets of $\mathcal{S}$. ||

Example 1.9 (Continuation of Example 1.6)**.** $\mathcal{F} = \{\phi,\, \mathcal{S},\, (4,\, 5),\, (4,\, 5)^c\}$ is a σ-field. ||

Remark 1.3. Note that there could be multiple σ-field on subsets of a sample space. Power set of sample space is always a σ-field and it is the largest σ-field. On the other hand $\{\mathcal{S},\, \emptyset\}$ is also a σ-field and it is the smallest σ-field. †

Definition 1.8 (Measurable Space)**.** *Let $\mathcal{S}$ be a sample space of a random experiment and $\mathcal{F}$ is a σ-field on subsets of $\mathcal{S}$. Then the ordered pair $(\mathcal{S},\, \mathcal{F})$ is called a measurable space.*

Definition 1.9 (Probability)**.** *Let $(\mathcal{S},\, \mathcal{F})$ be a measurable space. A set function $P : \mathcal{F} \to \mathbb{R}$ is called a probability if*

1. $P(E) \geq 0$ *for all* $E \in \mathcal{F}$.
2. $P(\mathcal{S}) = 1$.
3. *(Countable Additivity) Let $E_1,\, E_2,\, \ldots \in \mathcal{F}$ be a sequence of disjoint events (i.e., $E_i \cap E_j = \emptyset$ for all $i \neq j \in \mathbb{N}$) then*

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

The idea of probability germinates to predict the outcomes of gambling, where the classical definition of probability was used. When the people tried to give the axiomatic definition of the probability, it was observed that these three properties (mentioned in Definition 1.9) of the classical definition of probability are working fine. Hence, these three properties are used. Note that the first and the third axioms (mentioned in Definition 1.9) have good intuitions from the concepts of area of a region or volume of a shape. The area or volume is always non-negative and if we have several disjoint reasons or shapes, then the combined area or volume is the sum of the individual areas or volumes, respectively.

Definition 1.10 (Probability Space)**.** *Let $\mathcal{S}$ be a sample space and $\mathcal{F}$ be a σ-field on the subsets of $\mathcal{S}$. Let P be a probability defined on $\mathcal{F}$. The triplet $(\mathcal{S},\, \mathcal{F},\, P)$ is called a probability space.*

Example 1.10 (Continuation of Example 1.4)**.** Consider the random experiment of tossing of a coin, where sample space is $\mathcal{S} = \{H,\, T\}$ and $\mathcal{F} = \mathcal{P}\left(\mathcal{S}\right)$, the power set of $\mathcal{S}$. Consider a function $P : \mathcal{F} \to \mathbb{R}$ defined by

$$P(\mathcal{S}) = 1,\; P(\{H\}) = 0.6,\; P\left(\{T\}\right) = 0.4, \text{ and } P\left(\emptyset\right) = 0.$$

Here it is very easy to see that the first two axioms of Definition 1.9 are hold true. To check if the third axiom hold or not, let us consider the following cases. Note that here we have to choose E_i's such that E_i are disjoint.
Case I: $E_i = \emptyset$ for all $i \in \mathbb{N}$. Then $P\left(E_i\right) = 0$ for all $i \in \mathbb{N}$ implies $\sum_{i=1}^{\infty} P(E_i) = 0$. On the other hand, $P\left(\cup_{i=1}^{\infty} E_i\right) = P\left(\emptyset\right) = 0$.
Case II: $E_i = \mathcal{S}$ if $i = i_0$ for some $i_0 \in \mathbb{N}$ and $E_i = \emptyset$ for $i \neq i_0$. In this case $\sum_{i=1}^{\infty} P(E_i) = 1$ and $P\left(\cup_{i=1}^{\infty} E_i\right) = P\left(\mathcal{S}\right) = 1$.
Case III: $E_i = \{H\}$ if $i = i_0$ for some $i_0 \in \mathbb{N}$ and $E_i = \emptyset$ for $i \neq i_0$. In this case $\sum_{i=1}^{\infty} P\left(E_i\right) = P(\{H\}) = 0.6$ and $P\left(\cup_{i=1}^{\infty} E_i\right) = P\left(\{H\}\right) = 0.6$.
Case IV: $E_i = \{T\}$ if $i = i_0$ for some $i_0 \in \mathbb{N}$ and $E_i = \emptyset$ for $i \neq i_0$. In this case $\sum_{i=1}^{\infty} P\left(E_i\right) = P(\{T\}) = 0.4$ and $P\left(\cup_{i=1}^{\infty} E_i\right) = P\left(\{T\}\right) = 0.4$.
Case V: $E_i = \{T\}$ if $i = i_1$, $E_i = \{H\}$ if $i = i_2$ for some $i_1 \neq i_2 \in \mathbb{N}$, and $E_i = \emptyset$ for $i \neq i_1,\, i_2$. In this case $\sum_{i=1}^{\infty} P\left(E_i\right) = P\left(\{H\}\right) + P(\{T\}) = 1$ and $P\left(\cup_{i=1}^{\infty} E_i\right) = P\left(\mathcal{S}\right) = 1$. ||

Example 1.11. Consider a roll of a die. The sample space $\mathcal{S} = \{1,\, 2,\, \ldots,\, 6\}$ and take $\mathcal{F} = \mathcal{P}(\mathcal{S})$. Let $P(\emptyset) = 0$ and $P\left(i\right) = 1/6$ for $i \in \mathcal{S}$. Note that in this case the function $P(\cdot)$ have not defined for all the members in $\mathcal{F}$. However, if we assume that $P(\cdot)$ is a probability defined on the σ-field $\mathcal{F}$, we can uniquely extend $P(\cdot)$ for all other members of $\mathcal{F}$. Let $E \in \mathcal{F}$. As $\mathcal{S}$ is a finite set, so is E. Let the cardinality of E is n and the elements of E be $x_1 < x_2 < \ldots < x_n$. Define $E_i = \{x_i\}$ for $i = 1,\, 2,\, \ldots,\, n$ and $E_i = \emptyset$ for $i > n$. Clearly, E_i's are disjoint and $E = \cup_{i=1}^{\infty} E_i$. Now, if $P(\cdot)$ is a probability, using the third axiom of Definition 1.9, $P(E) = P\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = n/6$. ||

Example 1.12. Consider a roll of a die. The sample space $\mathcal{S} = \{1,\, 2,\, \ldots,\, 6\}$ and take $\mathcal{F} = \mathcal{P}(\mathcal{S})$. Let $P(\emptyset) = 0$ and $P\left(i\right) = i/21$ for $i \in \mathcal{S}$. As before, in this case also we can extend the function $P(\cdot)$ on $\mathcal{F}$ such that it becomes a probability on $\mathcal{F}$. ||

We have already pointed out that $\mathcal{P}(\mathcal{S})$ is a σ-field. Now, a natural question is that if $\mathcal{P}(\mathcal{S})$ is σ-field, then why do we define σ-field? Why should not we work with power set of sample space, always, and define probability on the power set of the sample space? We will try to answer these questions using next two examples. We will show that the choice of σ-field is an important issue.

Example 1.13. Let $\mathcal{S} = \{1,\, 2,\, \ldots,\, 60\}$ and $\mathcal{F} = \mathcal{P}(\mathcal{S})$. Let us define $P(E) = \frac{\#E}{\#\mathcal{S}}$ for all $E \in \mathcal{F}$. Note that as $\mathcal{S}$ is a finite set, $P(\cdot)$ satisfies all the axioms of probability. ||

Example 1.14. Now, consider a different problem where $\mathcal{S} = \mathbb{N}$ and $\mathcal{F} = \mathcal{P}(\mathbb{N})$. Can we extend the definition of probability in the previous example to define a probability for this example? A natural extension is

$$P(E) = \limsup_{n\to\infty} \frac{N_n(E)}{n}, \tag{1.1}$$

where $E \in \mathcal{F}$ and $N_n(E)$ is the number of times E occurs in the first n natural numbers. Here we have used lim sup instead of lim to overcome the issue of existence of limit of $\frac{N_n(E)}{n}$. Before answering if $P(\cdot)$ defined above is a probability, let us see the values of $P(\cdot)$ evaluated on some specified subsets of $\mathcal{S}$. Let us consider $A = \{\omega \in \mathbb{N} : \omega \text{ is a multiple of } 3\}$ and we want to calculate $P(A)$. Note that $N_n(A)$ is the number of multiple of three in the set $J_n = \{1, 2, \ldots, n\}$. Thus,

$$\frac{N_n(A)}{n} = \begin{cases} \frac{m}{3m} & \text{if } n = 3m \\ \frac{m}{3m+1} & \text{if } n = 3m+1 \\ \frac{m}{3m+2} & \text{if } n = 3m+2. \end{cases}$$

Hence, for all $n \in \mathbb{N}$, $\dfrac{1}{3 + \frac{6}{n-2}} \leq \dfrac{N_n(A)}{n} \leq \dfrac{1}{3}$ which implies $P(A) = \frac{1}{3}$. Similarly, $P(B) = \dfrac{1}{4}$ *(why?)* for $B = \{\omega \in \mathbb{N} : \omega \text{ is a multiple of } 4\}$. Now, assume that $C = \{2\}$. Then

$$\frac{N_n(C)}{n} = \begin{cases} 0 & \text{if } n = 1 \\ \frac{1}{n} & \text{if } n \geq 2. \end{cases}$$

Hence, $P(C) = 0$. Similarly, $P(D) = 0$ for any singleton set D. However, $\mathcal{S} = \mathbb{N} = \cup_{i\in\mathbb{N}}\{i\}$. Hence, if P satisfies the third axiom then $P(\mathcal{S}) = \sum_{i=1}^{\infty} P(\{i\}) = 0 \neq 1$, which contradicts the second axiom. Though $P(\cdot)$ as defined in (1.1) gives meaningful values for some sets like A and B, it does not satisfy all the three axioms, when it is defined on the power set of $\mathcal{S}$. ||

Note that we can always define a probability on the power set of a sample space. For example, let $\omega_0 \in \mathcal{S}$ be a fixed element. Define $P : \mathcal{P}(\mathcal{S}) \to \mathbb{R}$ by

$$P(A) = \begin{cases} 1 & \text{if } \omega_0 \in A \\ 0 & \text{if } \omega_0 \notin A. \end{cases}$$

It is easy to see that $P(\cdot)$ is a probability. However, in practice, a probability is used to model a practical situation, where the probability may need to satisfy extra conditions other then three conditions mentioned in Definition 1.9. The previous example suggests, depending on our objective we may need to choose from the set of all subsets of $\mathcal{S}$, certain subsets (not all) of $\mathcal{S}$ on which to define a probability P. For example, $P(\cdot)$ defined in (1.1) becomes a probability on the σ-fields $\mathcal{F}_1 = \{\mathcal{S}, \emptyset, A, A^c\}$ or $\mathcal{F}_2 = \{\mathcal{S}, \emptyset, A, B, A^c, B^c, A\cap B, A^c\cap B, A\cap B^c, A^c\cap B^c, A\cup B, A\cup B^c, A^c\cup B, A^c\cup B^c, (A\cap B)\cap(A^c\cap B^c), (A^c\cap B)\cup(A\cap B^c)\}$, where A and B are as defined in the previous example.

Next we will see some of the properties of probability. Let us assume that $(\mathcal{S}, \mathcal{F}, P)$ be a probability space.

Theorem 1.5. $P(\emptyset) = 0$.

Proof: Consider $E_i = \emptyset$ for all $i \in \mathbb{N}$. Clearly, E_i's are disjoint and $\cup_{i=1}^{\infty} E_i = \emptyset$. Using the third axiom of Definition 1.9,

$$P(\emptyset) + P(\emptyset) + \ldots = P(\emptyset) \implies P(\emptyset) = 0,$$

as using first axiom $P(\emptyset) \geq 0$. □

Theorem 1.6 (Finite Additivity). *If E_1, E_2, ..., E_n are n disjoint events, then*

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

Proof: Take $E_i = \emptyset$ for $i > n$ in the third axiom, to get the required result. □

Theorem 1.7 (Mononbooonicity). *$P(\cdot)$ is monotone, i.e., for E_1, $E_2 \in \mathcal{F}$ and $E_1 \subset E_2$, $P(E_1) \leq P(E_2)$.*

Proof: Note that $E_2 = (E_2 \cap E_1) \cup (E_2 \cap E_1^c)$ with $(E_2 \cap E_1) \cap (E_2 \cap E_1^c) = \emptyset$. Hence, using finite additivity, $P(E_2) = P(E_2 \cap E_1) + P(E_2 \cap E_1^c) = P(E_1) + P(E_2 \cap E_1) \geq P(E_1)$. Here the second equality is true as $E_1 \subset E_2$ and the last inequality is true as $P(\cdot) \geq 0$. □

The first and second terms in the right hand side of the decomposition $E_2 = (E_2 \cap E_1) \cup (E_2 \cap E_2^c)$ can be interpreted as E_2 occurring with E_1 and E_2 occurring without E_1, respectively. This decomposition is quite useful. We can use it to solve several problems in this course.

Theorem 1.8. *Let A, $B \in \mathcal{F}$ such that $P(B) = 0$. Then $P\left(A \cap B\right) = 0$.*

Proof: Note that $0 = P(B) \geq P\left(A \cap B\right) \geq 0$. Hence $P\left(A \cap B\right) = 0$. □

Theorem 1.9 (Subtractive Property). *$P(\cdot)$ is subtractive, i.e., for E_1, $E_2 \in \mathcal{F}$ and $E_1 \subset E_2$, $P(E_2 \setminus E_1) = P(E_2) - P(E_1)$.*

Proof: As in the proof of Theorem 1.7,

$$P(E_2) = P(E_1) + P(E_2 \cap E_1^c) \implies P(E_2 \setminus E_2) = P(E_2) - P(E_1).$$

□

Theorem 1.10. $0 \leq P(E) \leq 1$ *for all $E \in \mathcal{F}$.*

Proof: For any $E \in \mathcal{F}$, $\emptyset \subset E \subset \mathcal{S} \implies 0 \leq P(E) \leq 1$, using Theorem 1.7. □

Theorem 1.11. *If E_1, $E_2 \in \mathcal{F}$, then $P\left(E_1 \cup E_2\right) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.*

Proof: Note that $E_1 \cup E_2 = E_2 \cup (E_1 \setminus E_2)$. Also, $E_2 \cap (E_1 \setminus E_2) = \emptyset$. Hence, $P(E_1 \cup E_2) = P(E_2) + P\left(E_1 \setminus E_2\right)$. We have already seen that $P(E_1) = P(E_1 \cap E_2) + P(E_1 \setminus E_2)$. Combining, we get the required result. □

Theorem 1.12. *If E_1, $E_2 \in \mathcal{F}$, then $P\left(E_1 \cup E_2\right) \leq P(E_1) + P(E_2)$.*

Proof: Trivial from the last theorem. □

Theorem 1.13. *If $E \in \mathcal{F}$, then $P(E^c) = 1 - P(E)$.*

Proof: This is trivial as $\mathcal{S} = E \cup E^c$. □

Theorem 1.14. *Let A, $B \in \mathcal{F}$ with $P(B) = 1$. Then $P(A \cap B) = P(A)$.*

Proof: As $P(B^c) = 0$, $P(A \cap B^c) = 0 \implies P(A \cap B) = P(A)$. □

Definition 1.11 (Elementary Event)**.** *A single-ton event is called an elementary event.*

If $\mathcal{S}$ is finite, and $\mathcal{F} = \mathcal{P}(\mathcal{S})$, it is sufficient to assign probability to each elementary event in the sense that for any subset E of $\mathcal{S}$, we can calculate $P(E)$. For any $E \in \mathcal{F}$, E can be written as the union of elementary events that are in E, *i.e.*, $E = \cup_{\omega \in E} \{\omega\}$. Note that as E is finite (being a subset of $\mathcal{S}$, which is finite), there are finite number of elementary events in the expression. Also, elementary events are disjoint. Hence, $P(E) = \sum_{\omega \in E} P(\{\omega\})$.

Let $\mathcal{S}$ be finite, $\mathcal{F} = \mathcal{P}(\mathcal{S})$ and the elementary events be equally likely (*i.e.*, all the elementary events have same probability). Let the cardinality of the set $\mathcal{S}$ is n. Then $\mathcal{S}$ can be written as $\{\omega_1, \omega_2, \ldots, \omega_n\}$. Let $P(\{\omega_i\}) = c$ for all $i = 1, 2, \ldots, n$. Note that $\mathcal{S} = \cup_{i=1}^n \{\omega_i\}$ implies that $c = 1/n$. Now, for any event E, $P(E) = \frac{\#E}{n}$, which is classical probability. Hence, classical definition of probability is a particular case of the axiomatic definition of probability.

If $\mathcal{S}$ is countably infinite, and $\mathcal{F} = \mathcal{P}(\mathcal{S})$, it is still sufficient to assign probability to each elementary event. For any $E \in \mathcal{F}$, E is atmost countable, which means that E can be expressed as countable union of elementary events. Therefore, $P(E) = \sum_{\omega \in E} P(\{\omega\})$. However, in this case one cannot assign equal probability to each elementary event without violating the second axiom in Definition 1.9. Revisit Examples 1.11 and 1.12.

If $\mathcal{S}$ is uncountable, and $\mathcal{F} = \mathcal{P}(\mathcal{S})$, one can not make an equally likely assignment of positive probabilities to each elementary event. We can prove this statement by contradiction. If possible, suppose that an equally likely assignment of positive probability can be done, *i.e.*, $P(\{\omega\}) = c > 0$ for all $\omega \in \mathcal{S}$. $\mathcal{S}$ can be written as union of elementary events. However there are uncountable elementary events, and hence it is an uncountable union of sets. Therefore, the third axiom of probability can not be used directly to conclude that $P(\mathcal{S}) > 1$. Now, note that there exists a countable subset E of $\mathcal{S}$. Clearly, $P(\mathcal{S}) \geq P(E) = \infty$. This is a contradiction to the second axiom of Definition 1.9. Hence, our assumption is wrong.

Indeed, for uncountable $\mathcal{S}$ and $\mathcal{F} = \mathcal{P}(\mathcal{S})$, one can not assign positive probability to each elementary event without violating the axiom $P(\mathcal{S}) = 1$. This statement can be proved, again, by contradiction. If possible, suppose that $P(\{\omega\}) > 0$ for all $\omega \in \mathcal{S}$. Let us define the sets $A_n = \left\{\omega \in \mathcal{S} : P(\{\omega\}) > \frac{1}{n}\right\}$ for $n = 1, 2, \ldots$. The claim is that A_n is finite set for all $n = 1, 2, \ldots$. If not, then A_n is either countably infinite or uncountable. In both the cases, $P(A_n)$ is infinite, which is a contradiction. Hence, A_n is finite. Now, note that $\mathcal{S} = \cup_{n=1}^\infty A_n$. As A_n are finite, $\mathcal{S}$ is atmost countable, which is a contradiction and therefore our assumption that $P(\{\omega\}) > 0$ for all $\omega \in \mathcal{S}$ is wrong.

1.1.4 Continuity of Probability

Note that a function $f : \mathbb{R} \to \mathbb{R}$ is said to be continuous at x_0 if for every real sequence $\{x_n\}_{n \geq 1}$ converging to x_0, the sequence $\{f(x_n)\}_{n \geq 1}$ converges to $f(x_0)$. If we want to extend this definition of continuity of a function $f : \mathbb{R} \to \mathbb{R}$ to probability, first the convergence of sequence of events need to be defined, as the argument of $P(\cdot)$ is an event. Here we will consider the limits of increasing and decreasing sequences of events.

Definition 1.12 (Increasing Sequence of Events)**.** *A sequence, $\{E_n\}_{n \geq 1}$, of events are said to be increasing if $E_n \subseteq E_{n+1}$ for all $n = 1, 2, \ldots$.*

Definition 1.13 (Decreasing Sequence of Events)**.** *A sequence, $\{E_n\}_{n\geq 1}$, of events are said to be decreasing if $E_{n+1} \subseteq E_n$ for all $n = 1, 2, \ldots$.*

Definition 1.14 (Limit of Increasing Sequence of Events)**.** *For an increasing sequence, $\{E_n\}_{n\geq 1}$, of events, the limit is defined by $\lim\limits_{n\to\infty} E_n = \bigcup\limits_{n=1}^{\infty} E_n$.*

Definition 1.15 (Limit of Decreasing Sequence of Events)**.** *For a decreasing sequence, $\{E_n\}_{n\geq 1}$, of events, the limit is defined by $\lim\limits_{n\to\infty} E_n = \bigcap\limits_{n=1}^{\infty} E_n$.*

Theorem 1.15 (Continuity from below)**.** *Let $\{E_n\}_{n\geq 1}$ be an increasing sequence of events, then*

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

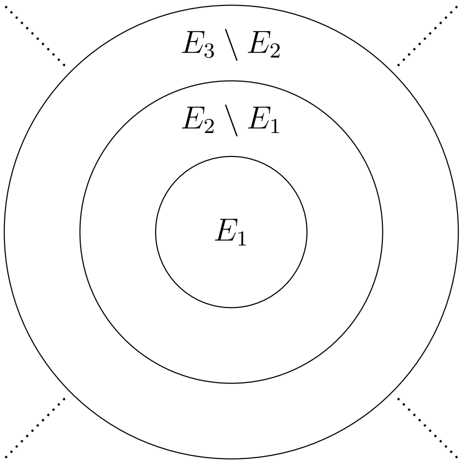


Figure 1.1: Sequence of events $\{A_n\}_{n\geq 1}$.

Proof: Let us define the following sequence, $\{A_n\}_{n\geq 1}$, of events as

$$A_1 = E_1 \text{ and } A_n = E_n \setminus E_{n-1} \text{ for } n = 2, 3, \ldots.$$

Please see the Figure 1.1. Clearly, A_n's are disjoint and $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} E_n$. Also,

$$P(A_n) = \begin{cases} P(E_1) & \text{if } n = 1 \\ P(E_n) - P(E_{n-1}) & \text{if } n = 2, 3, \ldots, \end{cases}$$

as $\{E_n\}_{n\geq 1}$ is an increasing sequence of events. Now,

$$P\left(\lim_{n\to\infty} E_n\right) = P\left(\bigcup_{n=1}^{\infty} E_n\right) = P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_i) = \lim_{N\to\infty} \sum_{n=1}^{N} P(A_n) = \lim_{n\to\infty} P(E_n).$$

Here the third equality is true for the third axiom of Definition 1.9, the fourth equality is true from the definition of a convergence of series, and the last equality is true for telescopic series. □

Theorem 1.16 (Continuity from above). *Let $\{E_n\}_{n\geq 1}$ be a decreasing sequence of events, then*

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

Proof: Let $A_n = E_n^c$ for all $n = 1, 2, \ldots$. Clearly, $\{A_n\}_{n\geq 1}$ is an increasing sequence of events. Hence,

$$\begin{aligned}
& P\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} P(A_n) \\
\implies & P\left(\bigcup_{n=1}^{\infty} E_n^c\right) = \lim_{n\to\infty} P\left(E_n^c\right) \\
\implies & P\left(\left(\bigcap_{n=1}^{\infty} E_n\right)^c\right) = \lim_{n\to\infty} \left(1 - P\left(E_n\right)\right) \\
\implies & P\left(\lim_{n\to\infty} E_n\right) = \lim_{n\to\infty} P(E_n).
\end{aligned}$$

□

1.2 Conditional Probability

We use conditional probability when we have some information about the outcome of a random experiment. Let us consider the following example.

Example 1.15. Let a die is thrown twice. Suppose that we are interested in the probability of the event that the sum of the outcomes of the rolls is six. Clearly, the sample space has 36 points and

$$\mathcal{S} = \{(n,\, m) : n,\, m = 1,\, 2, \ldots,\, 6\}.$$

As the sample space is finite, let us use the classical definition of probability. The required probability is 5/36.

Now, assume that you have observed that the first throw results in a 4. We are interested in the probability of the same event as before, but now we have extra information that the outcome of the first roll is 4. Note that when we know that the first roll results in a 4, the sample space changes and the new sample space is

$$\mathcal{S}_1 = \{(4,\, m) : m = 1,\, 2,\, \ldots,\, 6\},$$

which is the event that the first throw is a 4. We need to find the probability of the event that the sum is 6 in the sample space $\mathcal{S}_1$. There is only one case (4, 2) (in $\mathcal{S}_1$) which is favorable to the event of interest and hence the required probability is

$$\frac{1}{6} = \frac{1/36}{6/36} = \frac{P\left(A \cap H\right)}{P(H)},$$

where A and H are the events that sum is 6 and first roll results in 4, respectively. ||

Once you are given some information or you observe something, the sample space changes. Conditional probability is a probability on the changed sample space. Motivated by the above example, the definition of conditional probability is given as follows.

Definition 1.16 (Conditional Probability)**.** *Let H be an event with $P(H) > 0$. For any arbitrary event A, the conditional probability of A given H is denoted by $P(A|H)$ and defined by*

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

Note that to define the conditional probability, the probability of the conditioning event has to be positive. The probability of the intersection of two events can be expressed in terms of the conditional probability and the relationship is given below.

$$P(A \cap B) = \begin{cases} P(A)P(B|A) & \text{if } P(A) > 0 \\ P(B)P(A|B) & \text{if } P(B) > 0. \end{cases}$$

Definition 1.17 (Mutually Exclusive Events)**.** *A collection of events $\{E_1, E_2 \ldots\}$ is said to be mutually exclusive if $E_i \cap E_j = \emptyset$ for all $i \neq j$.*

Definition 1.18 (Exhaustive Events)**.** *A collection of events $\{E_1, E_2 \ldots\}$ is said to be exhaustive if $P\left(\cup_{i=1}^{\infty} E_i\right) = 1$.*

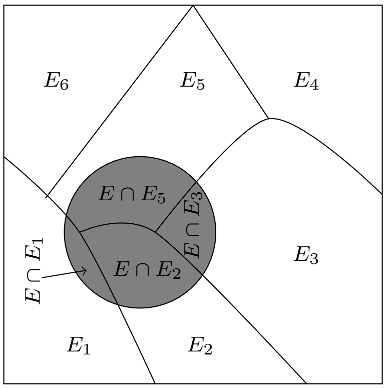
Thus if a collection of events $\{E_n\}_{n \geq 1}$ is such that $\cup_{n=1}^{\infty} E_n = \mathcal{S}$, then the collection is exhaustive.

Theorem 1.17 (Theorem of Total Probability)**.** *Let $\{E_1, E_2 \ldots\}$ be a collection of mutually exclusive and exhaustive events with $P(E_i) > 0$ for all $i = 1, 2, \ldots$. Then for any event E,*

$$P(E) = \sum_{i=1}^{\infty} P(E|E_i)P(E_i)\,.$$

Proof: Let us denote $\tilde{E}_i = E_i \cap E$ for all $i = 1, 2, \ldots$. Then $\left\{\tilde{E}_1, \tilde{E}_2, \ldots\right\}$ is mutually exclusive. Now, as E_i's are exhaustive, using Theorem 1.14

$$\begin{aligned} P(E) &= P\left(E \cap \left(\bigcup_{i=1}^{\infty} E_i\right)\right) \\ &= P\left(\bigcup_{i=1}^{\infty} (E \cap E_i)\right) \\ &= \sum_{i=1}^{\infty} P(E \cap E_i)\,, \text{ as } \tilde{E}_i \text{ are mutually exclusive} \\ &= \sum_{i=1}^{\infty} P(E|E_i)P(E_i). \text{ as } P(E_i) > 0 \text{ for all } i \in \mathbb{N}. \end{aligned}$$



□

The theorem of total probability tells that probability of an event can be computed by computing the probability of several partitions of the event. See the above figure, where the square and the shaded region indicates the sample space and the event E, respectively. In the figure $\{E_1, E_2, E_3, E_4, E_5, E_6\}$ is mutually exclusive and exhaustive. The event E can be partitioned into $E \cap E_1$, $E \cap E_2$, $E \cap E_3$, and $E \cap E_5$ and hence the probability of E can be computed by computing the probability of several partitions and then adding these probabilities.

Theorem 1.18 (Bayes' Theorem). *Let $\{E_1, E_2 \ldots\}$ be a collection of mutually exclusive and exhaustive events with $P(E_i) > 0$ for all $i = 1, 2, \ldots$. Let E be any event with $P(E) > 0$. Then*

$$P(E_i|E) = \frac{P(E|E_i)P(E_i)}{\sum_{j=1}^{\infty} P(E|E_j)P(E_j)} \quad \textit{for } i = 1, 2, \ldots.$$

Proof: Using the definition of conditional probability and the theorem of total probability, the proof is straight forward. □

In the theorem of total probability and Bayes' theorem, we have considered a countable collection of events $\{E_1, E_2, \ldots\}$. However, the theorems hold true even if we have a finite collection of mutually exclusive and exhaustive events *(Why?).*

Example 1.16. There are 3 boxes. Box 1 containing 1 white, 4 black balls. Box 2 containing 2 white, 1 black ball. Box 3 containing 3 white, 3 black balls. First you throw a fair die. If the outcomes are 1, 2 or 3 then box 1 is chosen, if the outcome is 4 then box 2 is chosen and if the outcome is 5 or 6 then box 3 is chosen. Finally, you draw a ball at random from the chosen box. Let W denote the event that the drawn ball is white. Also, assume that B_i, $i = 1, 2, 3$, denotes the event that ith box is selected after the roll of the die. Using Bayes' theorem, the (conditional) probability that the ball is from box 1 given the chosen ball is white is

$$P(B_1|W) = \frac{P(W|B_1)P(B_1)}{\sum_{i=1}^{3} P(W|B_i)P(B_i)} = \frac{9}{34}.$$

Similarly given the fact that the drawn ball is white, the probability that the ball is from box 2 is $P(B_2|W) = 5/17$. ||

1.3 Independence

Observe in the previous example that $P(B_1|W) = 9/34 < 1/2 = P(B_1)$, whereas $P(B_2|W) = 5/17 > 1/6 = P(B_2)$. Thus the occurrence of one event can make the occurrence of a second event more or less likely. Also, occurrence of an event may not change the probability of the occurrence of a second event. For example, let a coin is tossed two times. Then the probability of a head in the second toss does not change if the result of the first toss is a tail.

When occurrence of one event, say A, reduce the probability of the occurrence of another event, say B, we say that the events are negatively associated. That means A and B are negatively associated if $P(B|A) < P(B)$. For the conditional probability $P(B|A)$, $P(A)$ must be strictly greater than zero. Now, note that $P(B|A) < P(B)$ can be equivalently written as $P(A \cap B) < P(A)P(B)$, where we do not need the restriction $P(A) > 0$. Motivated by this discussion, we have the following definition.

Definition 1.19. *Let A and B be two events. They are said to be*

1. *negatively associated if $P(A \cap B) < P(A)P(B)$.*
2. *positively associated if $P(A \cap B) > P(A)P(B)$.*
3. *independent if $P(A \cap B) = P(A)P(B)$.*

Theorem 1.19. *If A and B are independent, so are A and B^c.*

Proof: As $A = (A \cap B) \cup (A \cap B^c)$, where $A \cap B$ and $A \cap B^c$ are disjoint. Hence

$$\begin{aligned} P\left(A \cap B^c\right) &= P(A) - P\left(A \cap B\right) \\ &= P(A) - P\left(A\right)P(B), \text{ as } A \text{ and } B \text{ are independent events} \\ &= P(A)P\left(B^c\right). \end{aligned}$$

Hence, A and B^c are independent events. □

Corollary 1.6. *If A and B are independent events, then*

1. *A^c and B are independent events.*

2. *A^c and B^c are independent events.*

Proof: This proof is simple using the previous theorem, and hence left as an exercise. □

Example 1.17. Let $P(B) = 0$. For any event A, $0 \leq P(A \cap B) \leq P(B) = 0$. Hence, $P(A \cap B) = 0$. On the other hand, $P(A)P(B) = 0$. Therefore, A and B are independent.

Now, assume that $P(B) = 1$. Then for any event A, A and B^c are independent as $P(B^c) = 0$. Using the previous theorem, A and B are independent events. In particular any event A is independent of $\mathcal{S}$ and $\emptyset$. ||

We have talked about independence of two events. A natural question is: Is the concept of independence be extended for more than two events? The answer is yes. However, there are two types of independence that are of interest for more than two events. We will discuss these concepts now.

Definition 1.20 (Pairwise Independent)**.** *A countable collection of events $E_1, E_2, \ldots$ are said to be pairwise independent if E_i and E_j are independent for all $i \neq j$.*

Definition 1.21 (Independent for Finite Collection of Events)**.** *A finite collection of events E_1, E_2, $\ldots$, E_n are said to be independent (or mutually independent) if for any sub-collection E_{n_1}, $\ldots$, E_{n_k} of $E_1, E_2, \ldots, E_n$,*

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

Definition 1.22 (Independent for Countable Collection of Events)**.** *A countable collection of events $E_1, E_2, \ldots$ are said to be independent (or mutually independent) if any finite sub-collection is independent.*

Suppose that we have three events E_1, E_2, and E_3 and we want to check if they are pairwise independent or not. We need to verify three conditions, *viz.*,

$$\begin{aligned} P(E_1 \cap E_2) &= P(E_1)P(E_2), \\ P(E_1 \cap E_3) &= P(E_1)P(E_3), \\ P(E_2 \cap E_3) &= P(E_2)P(E_3). \end{aligned}$$

However, to check if they are independent or not, we need to verify four conditions, *viz.*,

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

$$\begin{aligned}&P(E_1 \cap E_3) = P(E_1)P(E_3),\\&P(E_2 \cap E_3) = P(E_2)P(E_3),\\&P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2)P(E_3).\end{aligned}$$

That means to check if three events are independent or not, we need to check one extra condition over the conditions that need to verify for pairwise independence.

In general, one needs to verify $\binom{n}{2}$ conditions to check if a collection of n events are pairwise independent or not. To check if a collection of n events are independent or not, 2^n-n-1 conditions need to be verified. Clearly, if a collection of events are independent, then they are pairwise independent. However, in general, the converse is not true as illustrated by the following example.

Example 1.18. Suppose that a coin is tossed twice. The sample space has four points and is given by $\mathcal{S} = \{HH, HT, TH, TT\}$. Suppose that all elementary events are equally likely. That is $P(HH) = P(HT) = P(TH) = P(TT) = 1/4$. Let $E_1 = \{HH, HT\}$, $E_2 = \{HH, TH\}$ and $E_3 = \{HH, TT\}$. Clearly, E_1, E_2, and E_3 are the events that the first toss results in a heads, second toss results in heads, and both the tosses have same outcomes, respectively. It is easy to see that $P(E_1) = P(E_2) = P(E_3) = 1/2$. Also

$$\begin{aligned}P(E_1 \cap E_2) &= P(HH) = \frac{1}{4}.\\P(E_1 \cap E_3) &= P(HH) = \frac{1}{4}.\\P(E_2 \cap E_3) &= P(HH) = \frac{1}{4}.\\P(E_1 \cap E_2 \cap E_3) &= P(HH) = \frac{1}{4}.\end{aligned}$$

Thus $P(E_1 \cap E_2) = 1/4 = P(E_1)P(E_2)$, $P(E_1 \cap E_3) = 1/4 = P(E_1)P(E_3)$, and $P(E_2 \cap E_3) = 1/4 = P(E_2)P(E_3)$. This shows that the events E_1, E_2, and E_3 are pairwise independent. However, $P(E_1 \cap E_2 \cap E_3) = 1/4 \neq 1/8 = P(E_1)P(E_2)P(E_3)$. Hence, E_1, E_2, and E_3 are not independent. ||

Example 1.19. Let a die be rolled twice. The sample space is given by

$$\mathcal{S} = \{(i, j) : i = 1, \ldots, 6, \; j = 1, \ldots, 6\}.$$

Suppose all elementary events are equally likely, *i.e.*, $P(\omega) = 1/36$ for all $\omega \in \mathcal{S}$. Let us consider following events

$$\begin{aligned}E_1 &= \text{1st roll is 1, 2 or 3,}\\E_2 &= \text{1st roll is 3, 4 or 5,}\\E_3 &= \text{Sum of the rolls is 9.}\end{aligned}$$

Clearly, $P(E_1) = 1/2$, $P(E_2) = 1/2$, and $P(E_3) = 1/9$. Also, $P(E_1 \cap E_2 \cap E_3) = 1/36 = P(E_1)P(E_2)P(E_3)$. However, E_1 and E_2 are not independent as $P(E_1 \cap E_2) = 1/6 \neq 1/4 = P(E_1)P(E_2)$. Thus E_1, E_2, and E_3 are not independent, not even pairwise independent. This example shows that verifying the condition $P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2)P(E_3)$ is not enough to check if three events are independent or not. ||

Definition 1.23 (Conditional Independent)**.** *Given an event C two events A and B are said to be conditionally independent if $P(A \cap B|C) = P(A|C)P(B|C)$.*

Example 1.20. A box contains two coins: a fair regular coin and one fake two-headed coin (*i.e.*, $P(H) = 1$). The regular coin is called Coin 1 and the other is called Coin 2. You choose a coin at random and toss it twice. Define the following events.

$$\begin{aligned} A &= \text{First coin toss results in a } H. \\ B &= \text{Second coin toss results in a } H. \\ C &= \text{Coin 1 (regular) has been selected.} \end{aligned}$$

Here $P(A|C) = 1/2 = P(B|C)$, $P(A \cap B|C) = 1/4$. Hence, A and B are conditionally independent given C. As $P(A) = 3/4 = P(B)$ and $P(A \cap B) = 5/8$, A and B are not independent. Thus, the conditional independence does not imply independence in general. ||

Chapter 2

Random Variable

2.1 Random Variable

In most of the practical situations we are interested in numerical characteristic of a random experiment. For example, we may be interested in number of heads out of 10 tosses of a coin, number of bug reported for a newly developed software, value of total yield of a crop in different months of a year in Assam, the level of water in Brahmaputra river each day at a particular site, etc. Hence, it is helpful to use a function which maps a sample space to $\mathbb{R}$. Such a function is called a random variable. Moreover, we have rich mathematical tools on the set of real numbers. These mathematical tools can be used to analysis several properties of probability of the quantity of interest if we can transform any arbitrary sample space to $\mathbb{R}$ or a subset of $\mathbb{R}$.

Definition 2.1 (Random Variable)**.** *Let $\mathcal{S}$ be a sample space of a random experiment. A function $X : \mathcal{S} \to \mathbb{R}$ is called a random variable (RV).*

Example 2.1. Assume that a coin is tossed n times and the tosses are independent. Let $\mathcal{S}$ be the sample space for this random experiment. Let $X : \mathcal{S} \to \mathbb{R}$ be defined by the number of tails out of n tosses. Here, X is a RV, which can take values from the set $\{0, 1, \ldots, n\}$. We can compute probabilities that the RV takes several values from its' range, when we know the probability space $(\mathcal{S}, \mathcal{F}, P)$. For simplicity, suppose that $n = 2$. Then X takes value zero if and only if both the tosses result in heads. X takes value two if and only if both the tosses result in tails. X takes value one if and only if one of the tosses results in head and another in tail. Hence,

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

Note the technique of the computation of probabilities that the RV takes several values. To compute $P(X = x)$, we first find the inverse image of $X = x$ and then compute the required probability. ||

Example 2.2. Consider the random experiment of rolling a fair die twice. Assume that the throws are independent. Let $X : \mathcal{S} \to \mathbb{R}$ be defined by the sum of the outcomes of two rolls. Clearly, X is a RV. In this case, using the technique of the previous example, $P(X = 2) = 1/36, P(X = 3) = 2/36, P(X = 4) = 3/36, P(X = 5) = 4/36, P(X = 6) = 5/36, P(X = 7) = 6/36,$ $P(X = 8) = 5/36, P(X = 9) = 4/36, P(X = 10) = 3/36,$ $P(X = 11) = 2/36, P(X = 12) = 1/36.$ ||

Example 2.3. Suppose we are testing the reliability of a battery. In a reliability testing, an experimenter wants to know different characteristic of lifetime of a product. For example, one may want to know the average lifetime of batteries manufactured by a company, or proportion of batteries that can work beyond two years of use. In a typical reliability experiment, certain number of items are put on a life testing experiment and the failure times of the items are recorded. The outcomes are the lifetime of the product. Thus, the sample space can be taken as $\mathcal{S} = (0, \infty)$. Let us define $X_1 : \mathcal{S} \to \mathbb{R}$ by $X_1(\omega) = \omega$. Clearly, X_1 denote the lifetime of the battery. Now, suppose we are mainly interested in whether the battery would last more than 2 years or not. Then we can take a RV $X_2 : \mathcal{S} \to \mathbb{R}$ defined by $X_2(\omega) = I_{(2,\infty)}(\omega)$, where I_A denotes the indicator function of the set A and is defined by

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases} \tag{2.1}$$

Clearly, X_2 indicates if a battery has lifetime more than two years or not. If the lifetime of a battery is less than or equal to two years, then the value of X_2 is zero. On the other hand, the value of X_2 is one if the lifetime is more than two years.

Now, for some interval $I \subset (0, \infty)$, $P(I) = \int_I e^{-t}dt$ defines a probability on the Borel σ-field on the positive part of real line. This Borel σ-field is denoted by $\mathcal{B}(0, \infty)$. The exact definition of $\mathcal{B}(0, \infty)$ is not possible to provide in this course. However, all kind of intervals that are subset of $(0, \infty)$, single-ton sets that are subsets of $(0, \infty)$ belong to $\mathcal{B}(0, \infty)$. Though it is very difficult to construct, but there exist subsets of $(0, \infty)$ which does not belong to $\mathcal{B}(0, \infty)$. Thus, $\mathcal{B}(0, \infty)$ is not the power set of $(0, \infty)$. The Borel σ-field is a very standard and meaningful σ-field, which is used for almost all practical situations. Also, as the exact definition of Borel σ-algebra is not possible to give here, it is also not possible to prove that $P(I)$ defines a probability on $\mathcal{B}(0, \infty)$. Please take it as an information and proceed. We can, now, calculate the probability that $X_1 \leq x$ as

$$P(X_1 \leq x) = \int_0^x e^{-t}dt = 1 - e^{-x}$$

for $x > 0$. Also, we can calculate the probability of $X_2 = 0$ and $X_2 = 1$ as follows.

$$P(X_2 = 0) = P(X_1 \leq 2) = 1 - e^{-2} \quad \text{and} \quad P(X_2 = 1) = P(X_1 > 2) = 1 - P(X_1 \leq 2) = e^{-2}.$$

||

Definition 2.2 (Cumulative Distribution Function). *The cumulative distribution function (CDF) of a RV X is a function $F_X : \mathbb{R} \to [0, \infty)$ defined by*

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

Note that CDF is defined for all real numbers. Though in the definition $[0, \infty)$ is written as co-domain for CDF, it clear that CDF lies in the interval $[0, 1]$ for all real numbers.

Example 2.4 (Continuation of Example 2.1). As the random variable cannot take any negative value, $F_X(x) = P(X \leq x) = 0$ for all $x < 0$. Now, if we try to compute CDF at 0.5, we need to compute $P(X \leq 0.5) = P(X = 0) = 1/4$. The same argument will hold for all $x \in [0, 1)$, and hence, $F_X(x) = 1/4$ for all $x \in [0, 1)$.

Now, suppose we want to compute the CDF at 1.4. To compute it, we need to find $F_X(1.4) = P(X \leq 1.4) = P(X = 0 \text{ or } 1) = P(HH, HT, TH) = 1/4 + 1/2 = 3/4$. Just like the previous case, we can see that $F_X(x) = 3/4$ for all $x \in [1, 2)$. Proceeding in this way, we obtain the CDF of X as

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

||

Example 2.5 (Continuation of Example 2.2). Following the arguments of the last example, one can find the CDF of X and it is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 2 \\ 1/36 & \text{if } 2 \leq x < 3 \\ 3/36 & \text{if } 3 \leq x < 4 \\ 6/36 & \text{if } 4 \leq x < 5 \\ 10/36 & \text{if } 5 \leq x < 6 \\ 15/36 & \text{if } 6 \leq x < 7 \\ 21/36 & \text{if } 7 \leq x < 8 \\ 26/36 & \text{if } 8 \leq x < 9 \\ 30/36 & \text{if } 9 \leq x < 10 \\ 33/36 & \text{if } 10 \leq x < 11 \\ 35/36 & \text{if } 11 \leq x < 12 \\ 1 & \text{if } x \geq 12. \end{cases}$$

||

Example 2.6 (Continuation of Example 2.3). As lifetime cannot be negative, $P(X_1 \leq x) = 0$ for all $x < 0$. For $x > 0$, $P(X_1 \leq x) = P(0 < X_1 \leq x) = \int_0^x e^{-t} dt = 1 - e^{-x}$. Here, the first equality holds as $P(X < 0) = 0$. Hence, the CDF of X_1 is given by

$$F_{X_1}(x) = \begin{cases} 0 & \text{if} \quad x < 0, \\ 1 - e^{-x} & \text{if} \quad x \geq 0\,. \end{cases}$$

To find the CDF of X_2, note that X_2 can takes two values, *viz.*, 0 and 1. Hence, the CDF of X_2 is given by

$$F_{X_2}(x) = \begin{cases} 0 & \text{if} \quad x < 0, \\ 1 - e^{-2} & \text{if} \quad 0 \leq x < 1, \\ 1 & \text{if} \quad x \geq 1\,. \end{cases}$$

||

Theorem 2.1 (Properties of CDF). *The CDF of a RV has the following properties:*

1. $F_X(\cdot)$ *is non-decreasing.*

2. $\lim\limits_{x \uparrow \infty} F_X(x) = 1$.

3. $\lim\limits_{x \downarrow -\infty} F_X(x) = 0$.

4. $\lim\limits_{h \downarrow 0} F_X(x+h) = F_X(x)$ *for all* $x \in \mathbb{R}$. *Hence, CDF is right continuous.*

5. $\lim\limits_{h \downarrow 0} F_X(x-h) = F_X(x) - P(X = x)$ *for all* $x \in \mathbb{R}$.

Proof:

1. To show CDF is a non-decreasing function, we need to show that for $x_1 < x_2$, $F_X(x_1) \le F_X(x_2)$. Now, note that $\{X \le x_1\} \subset \{X \le x_2\}$, which implies $P(X \le x_1) \le P(X \le x_2)$. It proves the statement.

2. Let $\{x_n\}_{n \ge 1}$ be an increasing sequence of real numbers such that $x_n \to \infty$ as $n \to \infty$. Let us define a sequence of events $A_n = \{\omega \in \mathcal{S} : X(\omega) \in (-\infty, x_n]\}$ for all $n \ge 1$. As $\{x_n\}_{n \ge 1}$ is increasing, $\{A_n\}_{n \ge 1}$ is also increasing sequence of events. Thus,
$$\lim_{n \to \infty} A_n = \cup_{n=1}^{\infty} A_n = \mathcal{S}.$$
Now,
$$\lim_{n \to \infty} F_X(x_n) = \lim_{n \to \infty} P(A_n) = P\left(\lim_{n \to \infty} A_n\right) = P(\mathcal{S}) = 1.$$
This shows that for any increasing sequence $\{x_n\}_{n \ge 1}$ of real numbers with $x_n \to \infty$ as $n \to \infty$, $\lim_{n \to \infty} F_X(x_n) = 1$. Thus, $\lim_{x \to \infty} F_X(x) = 1$.

3. Let $\{x_n\}_{n \ge 1}$ be a decreasing sequence of real numbers such that $x_n \to -\infty$ as $n \to \infty$. Take $B_n = \{\omega \in \mathcal{S} : X(\omega) \le x_n\}$ for all $n = 1, 2, \ldots$. Then $\{B_n\}_{n \ge 1}$ is a decreasing sequence events and $\lim_{n \to \infty} B_n = \emptyset$. Hence,
$$\lim_{n \to \infty} F_X(x_n) = \lim_{n \to \infty} P(B_n) = P\left(\lim_{n \to \infty} B_n\right) = 0.$$
This shows that for any decreasing sequence $\{x_n\}_{n \ge 1}$ of real numbers such that $x_n \to -\infty$ as $n \to \infty$, $\lim_{n \to \infty} F_X(x_n) = 0$. Hence, $\lim_{x \to -\infty} F_X(x) = 0$.

4. Fix $x \in \mathbb{R}$. Let $\{x_n\}_{n \ge 1}$ be a decreasing sequence of real numbers such that $x_n \to x$ as $n \to \infty$. Assume that $C_n = \{\omega \in \mathcal{S} : X(\omega) \in (-\infty, x_n]\}$ for all $n \ge 1$. Clearly, $\{C_n\}_{n \ge 1}$ is decreasing sequence of events. Hence,
$$\lim_{n \to \infty} C_n = \cap_{n=1}^{\infty} C_n = \{\omega \in \mathcal{S} : X(\omega) \in (-\infty, x]\}.$$
Now,
$$\lim_{n \to \infty} F_X(x_n) = \lim_{n \to \infty} P(C_n) = P\left(\lim_{n \to \infty} C_n\right) = F_X(x).$$
This completes the proof.

5. Fix $x \in \mathbb{R}$. Taking an increasing sequence $\{x_n\}_{n \ge 1}$ of real numbers such that $x_n \to x$ as $n \to \infty$, we can prove this part like the previous parts. *(Complete it.)*

□

Theorem 2.2. *Let $G : \mathbb{R} \to \mathbb{R}$ be a function satisfying properties 1–4 of the Theorem 2.1. Then $G(\cdot)$ is a CDF of a RV.*

Proof: The proof of the theorem is out of the scope of this course. □

Though the proof is out of scope, it is an important theorem. This theorem can be use to check if a given function is a CDF or not. We need to check if Properties 1–4 of the Theorem 2.1 are satisfied by the given function or not. If the function satisfy all the four properties, then it is a CDF. Otherwise, it is not a CDF.

By distribution (or probability distribution) of a RV, we mean how the probability is distributed over the real line for the RV. One of the ways to see the distribution of a RV is through CDF of the RV. Note that a random variable is just a function defined on the sample space and does not depend on probability that is defined on the σ-field. However, the distribution of the RV depends on the probability. Hence, keeping the function same if we change the probability then the RV will remain same but its distribution will change. Consider the Example 2.1 with changed probabilities $P(HH) = 9/16, P(TT) = 1/16, P(HT) = P(TH) = 3/16$. The CDF of X in this case is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{9}{16} & \text{if } 0 \leq x < 1 \\ \frac{15}{16} & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2. \end{cases}$$

Definition 2.3 (Atom)**.** *If $x \in \mathbb{R}$ is such that $P(X = x) > 0$, then x is said to be an atom of the CDF of X or simple atom of X.*

The first property of a CDF says that a CDF is always non-decreasing. That means that a CDF can have only jump discontinuities. From the third property of a CDF, we know that a CDF is always right continuous. However, it can have a discontinuity from the left due to the fourth property. Thus, if a CDF has a discontinuity at a point x, then x is an atom of the CDF. On the other hand if the distribution function of a RV has no atoms, then the CDF of the RV is a continuous function.

We can write probabilities of different events relating to a random variables in terms of CDF of the random variable. Consider the following cases in this regard. By definition of the CDF, $P(X \leq a) = F_X(a)$. From the fifth property of CDF (Theorem 2.1), we have $P(X < a) = \lim_{x \uparrow a} F_X(x) = F_X(a-)$ and $P(X = a) = F_X(a) - F_X(a-)$. Moreover,

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a). \\ P(a \leq X \leq b) &= P(X \leq b) = P(X < a) = F_X(b) - F_X(a-). \\ P(a < X < b) &= P(X < b) - P(X \leq a) = F_X(b-) - F_X(a). \\ P(a \leq X < b) &= P(X < b) - P(X < a) = F_X(b-) - F_X(a-). \end{aligned}$$

2.2 Discrete Random Variable

Definition 2.4 (Discrete Random Variable)**.** *A RV is said to have discrete distribution if there exists an atmost countable set $S_X \subset \mathbb{R}$ such that $P(X = x) > 0$ for all $x \in S_X$ and $\sum_{x \in S_X} P(X = x) = 1$. S_X is called the support of X. A RV having discrete distribution is called a DRV.*

Definition 2.5 (Probability Mass Function)**.** *Let X be a RV having discrete distribution with support S_X. Define a function $f_X : \mathbb{R} \to [0, 1]$ by*

$$f_X(x) = P(X = x) = \begin{cases} P(X = x) & \text{if } x \in S_X \\ 0 & \text{otherwise.} \end{cases}$$

The function f_X is called the probability mass function (PMF) of X.

Example 2.7 (Continuation of Example 2.4)**.** In Example 2.4, we have seen the CDF of the RV X is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{4} & \text{if } 0 \le x < 1 \\ \frac{3}{4} & \text{if } 1 \le x < 2 \\ 1 & \text{if } x > 2. \end{cases}$$

Let us check if X is DRV. Let us consider the set $D = \{0,\, 1,\, 2\}$, which is a finite set, and hence, it is atmost countable. Now,

$$\begin{aligned} P\left(X = 0\right) &= F_X(0) - F_X(0-) = \frac{1}{4} > 0. \\ P\left(X = 1\right) &= F_X(1) - F_X(1-) = \frac{1}{2} > 0. \\ P\left(X = 2\right) &= F_X(2) - F_X(2-) = \frac{1}{4} > 0. \end{aligned}$$

Also, $P(X \in D) = P(X = 0) + P(X = 1) + P(X = 2) = 1$. In this case, we have an atmost countable set $S_X = D$, such that $P(X = x) > 0$ for all $x \in S_X$ and $\sum_{x \in S_X} P(X = x) = 1$. Hence, X is a DRV. ||

Note that $P(X = x)$ is strictly greater than zero if and only if there is an atom at x. In other words, $P(X = x) > 0$ if and only if x is a point of discontinuity of the CDF of the RV. Hence, to check if a RV is DRV or not, we should start with the set of all points of discontinuities of the CDF of the RV. If D denote the set of all points of discontinuities of $F_X(\cdot)$, then we should check if $P(X \in D)$ equal one or not. If $P(X \in D) = 1$, then the corresponding random variable is DRV. If not, then it is not a DRV.

Example 2.8 (Continuation of Example 2.5)**.** The RV X in Example 2.5 can be shown a DRV by taking $S_X = \{2,\, 3, \ldots,\, 12\}$. I leave it as a practice problem and please complete it. ||

Example 2.9 (Continuation of Example 2.6)**.** The CDF of the RVs X_1 is a continuous function, hence, there does not exists any atmost countable set S_X, such that $P\left(X \in S_X\right) = 1$. Therefore, X_1 is not a DRV. You can easily show that X_2 is a DRV. ||

Note that if we are given with a CDF, in principle we should be able to say whether the RV is discrete or not. If the RV is discrete, we should also be able to find the PMF of the RV using the formula

$$f_X(x) = F_X(x) - F_X(x-).$$

On the other hand, if a PMF of a DRV is given, the CDF of the RV can be computed as

$$F_X(x) = \sum_{\substack{y \in S_X \\ y \leq x}} f_X(y).$$

Thus, for a DRV, CDF and PMF have a one-one correspondence in the sense that if one of them is given, then other one can be found uniquely.

Theorem 2.3 (Properties of PMF). *Let X be a DRV with PMF $f_X(\cdot)$ and support S_X. Then*

1. *$f_X(x) \geq 0$ for all $x \in \mathbb{R}$.*
2. *$\sum\limits_{x \in S_X} f_X(x) = 1$.*

Proof: Straight forward using the definition of PMF. □

Theorem 2.4. *Suppose a real valued function $h : \mathbb{R} \to \mathbb{R}$ satisfies the following two conditions:*

1. *$h(x) \geq 0$ for all $x \in \mathbb{R}$ and $D = \{x : h(x) > 0\}$ is atmost countable.*
2. *$\sum_{x \in D} h(x) = 1$.*

Then $h(\cdot)$ is a probability mass function of some DRV.

Proof: The proof of this theorem is out of scope of this course. □

Example 2.10 (Bernoulli Distribution). Consider a random experiment which has two possible outcomes. Such a random experiment is called Bernoulli experiment or Bernoulli trial. Examples of Bernoulli experiment include tossing a coin, test if a person is infected with novel corona virus, checking if a mobile phone works more than three years, etc. It is customary to name one of the outcome as success and other as failure. Thus, in a coin toss experiment, we may say that getting a tail is a success and head is failure. In a COVID19 testing experiment, we may say having the virus is a success and not having it a failure. Suppose that the probability of success is $p \in [0,\, 1]$ and that of failure is $1-p$. In this case, the sample space is $\mathcal{S} = \{S,\, F\}$, where S and F denote a success and a failure, respectively. Let us define a RV

$$X = \begin{cases} 1 & \text{if a success occurs} \\ 0 & \text{if a failure occurs.} \end{cases}$$

Then it is clear that $P(X = 1) = p$ and $P(X = 0) = 1 - p$. The corresponding CDF is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

It is easy to show that X is a DRV with PMF

$$f_X(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1. \end{cases} \tag{2.2}$$

The distribution of a DRV having the PMF given in (2.2) is called a Bernoulli distribution with parameter $p \in [0,\, 1]$ and the RV is called a Bernoulli RV. We will use $X \sim \mathit{Bernoulli}(p)$ to denote that the RV X follows a Bernoulli distribution with parameter p. ||

Example 2.11 (Binomial Distribution). Consider a Bernoulli experiment with probability of success p. The experiment is repeated n times independently. The sample space is given by

$$\mathcal{S} = \{(\omega_1,\, \omega_2,\, \ldots,\, \omega_n) : \omega_i \in \{S,\, F\} \text{ for all } i = 1,\, 2,\, \ldots,\, n\}.$$

Let the RV X denote the number of successes that occur out of n trials of the Bernoulli experiment. Clearly, X takes values in the set $D = \{0,\, 1,\, \ldots,\, n\}$. Let us try to calculate $P\left(X = k\right)$ for $k \in D$. The event $X = k$ means that there are exactly k successes and $n-k$ failures. Now, the probability of getting a particular arrangement of k successes and $n-k$ failures is $p^k(1-p)^{n-k}$. The event $X = k$ will occurs if any one of $\binom{n}{k}$ arrangements of k successes and $n-k$ failures occurs. Hence,

$$P\left(X = k\right) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k \in D.$$

Using binomial expansion, $\sum_{k\in D} P\left(X = k\right) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$. Hence, X is a DRV with PMF

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x = 0,\, 1,\, \ldots,\, n \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

The distribution of a DRV having the PMF given in (2.3) is called a binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0,\, 1]$ and the RV is called a binomial RV. We will use $X \sim \mathit{Bin}(n,\, p)$ to denote that the RV X follows a binomial distribution with parameters n and p. ||

Example 2.12 (Geometric Distribution). Let a Bernoulli experiment with success probability p be repeated again and again, independently. The sample space is

$$\mathcal{S} = \{(\omega_1,\, \omega_2,\, \ldots) : \omega_i \in \{S,\, F\} \text{ for all } i = 1,\, 2,\, \ldots\}.$$

Let X be a function that denotes the number failures before the first success. Clearly, the range of X is $D = \{0,\, 1,\, \ldots\}$. Then for $k \in D$, $P(X = k) = p(1-p)^k$. To see it, notice that the event $X = k$ occurs if and only if there are k failures occur in first k trials, and then a success occurs at the $(k+1)$st trial. As the trials are independent, $P(X = k) = p(1-p)^k$. Using the geometric series, it is easy to see that

$$\sum_{k\in D} P(X = k) = \sum_{k=0}^{n} p(1-p)^k = 1.$$

Hence, X is a DRV with PMF

$$f_X(x) = \begin{cases} p(1-p)^x & \text{if } x = 0,\, 1,\, 2,\, \ldots \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

The distribution of a DRV having the PMF given in (2.4) is called a geometric distribution with parameter $p \in [0, 1]$ and the RV is called a geometric RV. We will use $X \sim Geo(p)$ to denote that the RV X follows a geometric distribution with parameter p. ||

Example 2.13 (Poisson Distribution)**.** Consider the function

$$f_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{if } x = 0,\, 1,\, 2,\, \ldots \\ 0 & \text{otherwise,} \end{cases} \tag{2.5}$$

where $\lambda > 0$. As $f_X(x) \geq 0$ for all $x \in \mathbb{R}$, $D = \{x \in \mathbb{R} : f_X(x) > 0\}$ is countable, and $\sum_{x=0}^{\infty} f_X(x) = 1$, $f_X(\cdot)$ is a PMF. The distribution of a DRV having the PMF given in (2.5) is called a Poisson distribution with parameter $\lambda > 0$ and the RV is called a Poisson RV. We will use $X \sim Poi(\lambda)$ to denote that the RV X follows a Poisson distribution with parameter λ.

Unlike the previous cases, we do not discuss the motivation of Poisson distribution here. We will comeback to the issue later. However, this is a very useful distribution, which has a wide range of applications in a diverse number of areas. Applications of Poisson distribution include modeling of number of accidents in a particular spot of a city, number of misprints in each page of a book having large number of pages, number of dinners in a certain restaurant, number of customers on a server at a given time, etc. ||

Example 2.14. Suppose that an airplane engine will fail, when in flight, with probability $1-p$ independently from engine to engine. The airplane will make a successful flight if at least 50 percent of its engines remain operating. As each engine is assumed to fail or function independently of what happens with the other engines, it follows that the number of engines remaining operative is a binomial RV with parameter p. Hence, the probability that a four-engine plane makes a successful flight is

$$\binom{4}{2}p^2(1-p)^2 + \binom{4}{3}p^3(1-p) + \binom{4}{4}p^4 = 6p^2(1-p)^2 + 4p^3(1-p) + p^4,$$

whereas the corresponding probability for a two-engine plane is

$$\binom{2}{1}p(1-p) + \binom{2}{2}p^2 = 2p(1-p) + p^2.$$

Hence, the four-engine plane is safer if

$$6p^2(1-p)^2 + 4p^3(1-p) + p^4 \geq 2p(1-p) + p^2 \implies p \geq \frac{2}{3}.$$

Therefore, the four-engine plane is safer when the engine success probability is at least as large as $\frac{2}{3}$, where the two-engine plane is safer if this probability falls below $\frac{2}{3}$. ||

2.3 Continuous Random Variable

Definition 2.6 (Continuous Random Variable and Probability Density Function)**.** *A RV is said to have a continuous distribution if there exists a non-negative integrable function* $f_X : \mathbb{R} \to [0, \infty)$ *such that*

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

for all $x \in \mathbb{R}$. A RV having a continuous distribution is called a continuous RV (CRV). The function f_X is called the probability density function (PDF). The set $S_X = \{x \in \mathbb{R} : f_X(x) > 0\}$ is called support of X.

Following the definition of CRV, there exists a PDF and $P(X \leq x)$ can be written as the area under the PDF from $-\infty$ to x. Note that for DRV we have PMF and PMF is only defined for DRV, where the PDF is only defined for CRV.

For a continuous RV X, $P(X = a) = 0$ for all $a \in \mathbb{R}$. To see it notice that $P(X = a)$ can be interpreted as the integration of PDF over a single-ton set $\{a\}$. As we know that the integration over a single-ton set is zero, $P(X = a) = 0$ for all $a \in \mathbb{R}$. That means the CDF of a CRV does not have any atom, and hence, the CDF of a CRV is continuous.

Let $f_X(\cdot)$ be a PDF of a CRV. Let a and $b \geq 0$ be real numbers such that $b \neq f_X(a)$. Define a new function as

$$g(x) = \begin{cases} f_X(x) & \text{if } x \neq a \\ b & \text{if } x = a. \end{cases}$$

It is easy to see that $g(x) \geq 0$ and $F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{x} g(t)dt$. Hence, $g(\cdot)$ is a PDF corresponding to the CDF $F_X(\cdot)$ and $g(x) \neq f_X(x)$. Thus, PDF is not unique. As a consequence, support of a CRV is also not unique. Note that PMF and support of a DRV are unique. Also, note that $f_X(x)$ is not $P(X = x)$ for $P(X = x)$ is always zero, but $f_X(x)$ can be greater than zero.

For a CRV with PDF $f_X(x)$,

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(t)dt - \int_{-\infty}^{a} f_X(t)dt = \int_{a}^{b} f_X(t)dt.$$

Also, note that $P(a \leq X \leq b) = F_X(b) - F_X(a-) = F_X(b) - F_X(a)$, as the CDF of a CRV is continuous. Therefore, for a CRV X,

$$P\left(a < X < b\right) = P\left(a \leq X < b\right) = P\left(a < X \leq b\right) = P\left(a \leq X \leq b\right) = \int_{a}^{b} f_X(t)dt.$$

Example 2.15 (Uniform Distribution)**.** Let the CDF of a RV X is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{if } x \geq b, \end{cases}$$

where $-\infty < a < b < \infty$. It is very easy to see that for all $x \in \mathbb{R}$, $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$, where

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

It is also clear that $f_X(x) \geq 0$ for all $x \in \mathbb{R}$. Hence, X is a CRV with PDF (2.6). The distribution of a CRV having the PDF given in (2.6) is called a uniform distribution with parameters a and b ($-\infty < a < b < \infty$) and the RV is called a uniform RV. We will use $X \sim U(a,\, b)$ to denote that the RV X follows a uniform distribution with parameters a and b.

Note that $f_X(x)$ is constant in the interval $(a,\, b)$ and X is just as likely to be near any value on $(a,\, b)$ as any other value. To check this, note that for any $a < \alpha < \beta < b$,

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} f_X(x)dx = \frac{\beta - \alpha}{b - a}.$$

In other words, the probability that X is in any particular subinterval of $(a,\, b)$ is proportional to the length of that subinterval. ||

Example 2.16 (Exponential Distribution)**.** Consider that X be a RV having CDF given by

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.7}$$

where $\lambda > 0$. Let us define

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

Clearly, for all $x \in \mathbb{R}$, $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ and $f_X(x) \geq 0$. Hence, X is a CRV with PDF given in (2.8). The distribution of a CRV having the PDF given in (2.8) is called an exponential distribution with parameter $\lambda > 0$ and the RV is called an exponential RV. We will use $X \sim Exp(\lambda)$ to denote that the RV X follows an exponential distribution with parameter λ. Plot of PDFs and CDFs are given in Figure 2.1 for different values of λ.

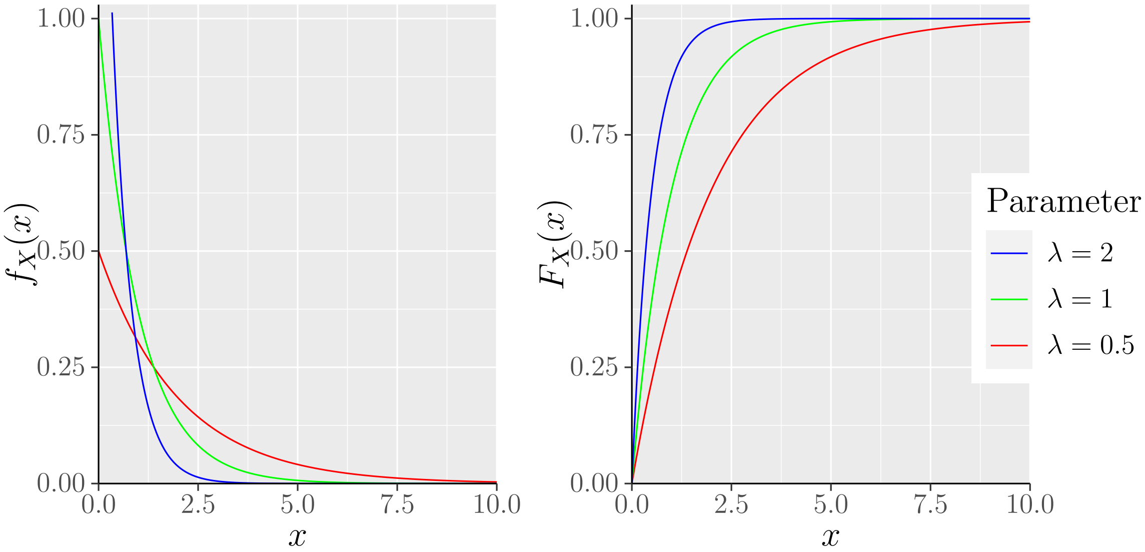


Figure 2.1: Plot of PDFs and CDFs of $Exp(\lambda)$ for several values of λ.

Clearly, PDFs are bounded and decreasing on the positive part of real line. The CDF rushes towards one more quickly for smaller values of λ compared to larger values of the parameter. This means that as λ increases, $P(X \leq x)$ decreases for each fixed $x > 0$. ||

Example 2.17 (Normal Distribution)**.** Let X be a RV with CDF

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \text{ if } -\infty < x < \infty,$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. It is clear that the corresponding PDF is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ if } -\infty < x < \infty, \tag{2.9}$$

The distribution of a CRV having the PDF given in (2.9) is called a normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2\,(\sigma > 0)$ and the RV is called a normal RV. We will use $X \sim N(\mu,\, \sigma^2)$ to denote that the RV X follows a normal distribution with parameters μ and σ^2. The normal distribution with $\mu = 0$ and $\sigma = 1$ is called a standard normal distribution. If $X \sim N(0,\, 1)$, then X is called standard normal RV. The CDF and PDF of a standard normal distribution are denoted by $\Phi(\cdot)$ and $\phi(\cdot)$, respectively. Plots of PDFs and CDFs are given in Figure 2.2 for different values of the parameters. The following points are quite relevant for the plot. The location of the PDF changes with change in μ keeping σ^2 fixed. On the other hand, keeping μ fixed, if we increase σ^2, the PDF becomes flatter and less peaked. Also, the PDF is symmetric with respect to the point μ. $||$

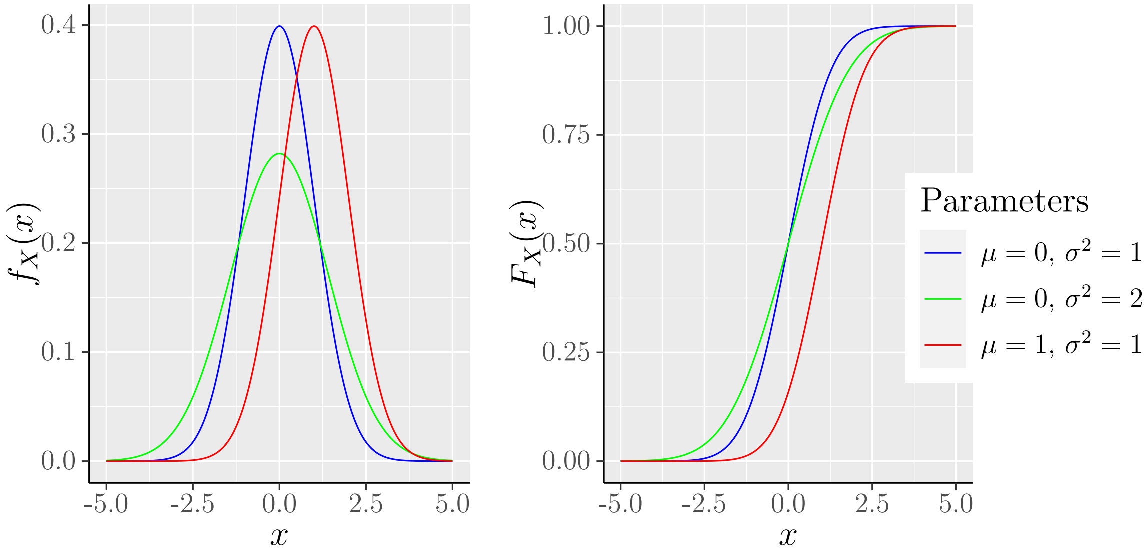


Figure 2.2: Plot of PDFs and CDFs of $N(\mu,\, \sigma^2)$ for several values of parameters.

Theorem 2.5 (Properties of PDF). *Let X be a CRV with PDF $f_X(\cdot)$. Then*

1. $f_X(x) \geq 0$ *for all* $x \in \mathbb{R}$.

2. $\int_{-\infty}^{\infty} f_X(x) = 1$.

Proof: 1. Straight forward form the definition of PDF.

2.

$$\int_{-\infty}^{\infty} f_X(x)dx = \lim_{A\to\infty} \int_{-\infty}^{A} f_X(x)dx = \lim_{A\to\infty} F_X(A) = 1.$$

□

Theorem 2.6. *Suppose a real valued function $g : \mathbb{R} \to \mathbb{R}$ satisfies the following conditions:*

1. $g(x) \geq 0$ *for all* $x \in \mathbb{R}$.

2. $\int_{-\infty}^{\infty} g(x)dx = 1$.

Then $g(\cdot)$ is a PDF of some CRV.

Proof: The proof of this theorem is out of scope of the course. □

Example 2.18 (Gamma Distribution)**.** It can be shown that the improper integral

$$\int_0^\infty t^{\alpha-1}e^{-t}dt$$

converges for all $\alpha > 0$ and diverges for all $\alpha \leq 0$. For the proof of the statement along with some properties of the integral, please see Appendix 2.A at the end of this chapter. For $\alpha > 0$ the integral is denoted by $\Gamma(\alpha)$ and is called gamma integral. Thus,

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt \quad \text{for } \alpha > 0.$$

Gamma integral gives a probability distribution of a CRV, which is described below. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

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

where $\alpha > 0$ and $\beta > 0$. It is easy to see that the function $f(\cdot)$ satisfies both the conditions of the previous theorem. Hence, $f(\cdot)$ is a PDF of a CRV. Thus, we can define Gamma distribution as follows. A RV X is said to have a gamma distribution if the PDF of the RV is given by

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

where $\alpha > 0$ and $\beta > 0$. In this case, the RV X is called a gamma RV. We will use the notation $X \sim Gamma(\alpha,\, \beta)$ to denote the fact that X follows a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. The plot of PDFs and CDFs of $Gamma(\alpha,\, \beta)$ are given in Figure 2.3. The following points are quite visible form the plot. On the positive part of real line, for $\alpha \leq 1$, the PDF is strictly decreasing, where the PDF has a unique maxima for $\alpha > 1$. The PDF is unbounded at the point $x = 0$ for $\alpha < 1$. As the shape of the PDF changes with a change in α keeping β fixed, α is called a shape parameter. If we change β keeping α fixed, the flatness of the PDF changes. Hence, β is called a scale parameter. ||

Example 2.19 (Beta Distribution)**.** The improper integral

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

converges if and only if $\alpha > 0$ and $\beta > 0$. For the proof of this statement and some main properties of the integral, please see the Appendix 2.B. For $\alpha > 0$ and $\beta > 0$, this improper integral is denoted by $B(\alpha,\, \beta)$ and is called beta integral. Thus,

$$B(\alpha,\, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \quad \text{for } \alpha > 0 \text{ and } \beta > 0.$$

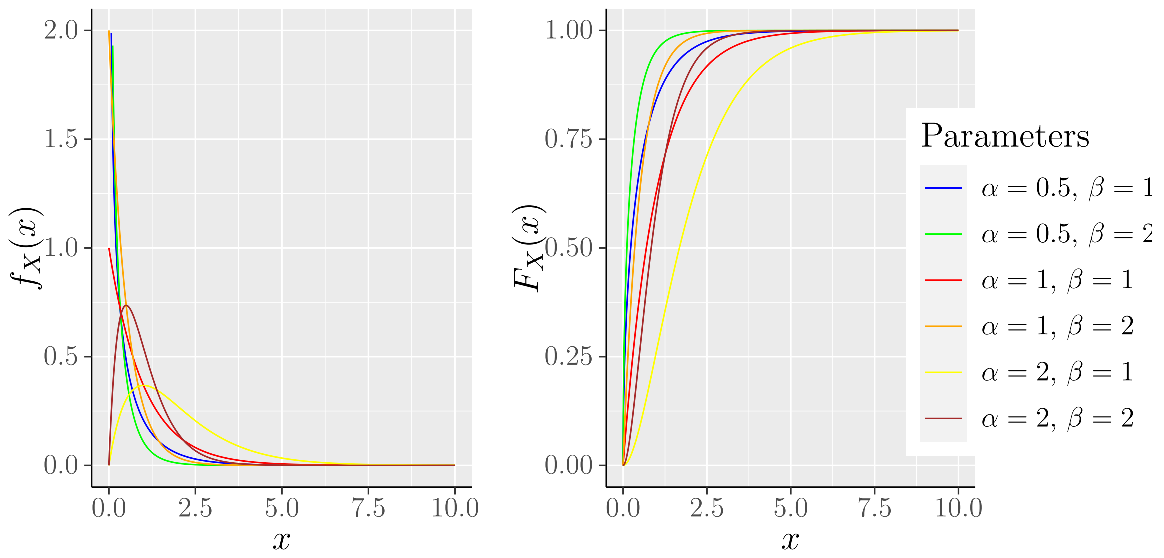


Figure 2.3: Plot of PDFs and CDFs of $Gamma(\alpha,\ \beta)$ for several values of parameters.

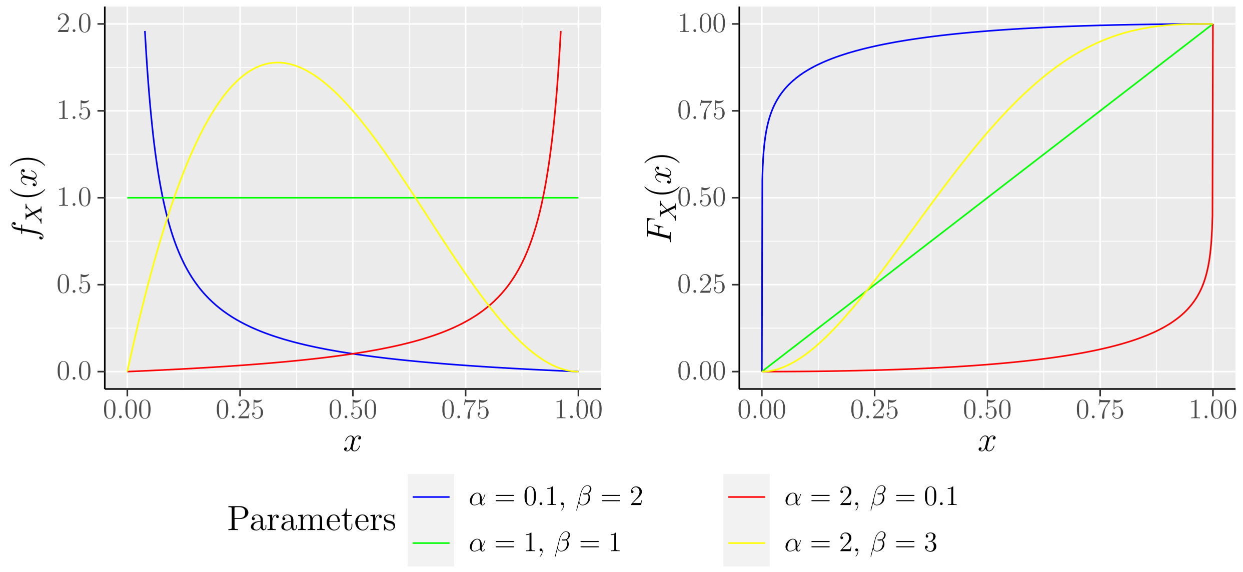


Figure 2.4: Plot of PDFs and CDFs of $Beta(\alpha,\ \beta)$ for several values of parameters.

Like gamma integral, beta integral also gives a PDF, and hence, a distribution of a CRV. Consider the function

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)}\, x^{\alpha-1}(1-x)^{\beta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $f(\cdot)$ is a PDF. Thus, we have the following definition of beta distribution. A RV X is said to have a beta distribution if the PDF of the RV is given by

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

where $\alpha > 0$ and $\beta > 0$. In this case, the RV X is called beta RV with parameters α and β. We will use the notation $X \sim Beta(\alpha,\, \beta)$ to denote the fact that the distribution of X is beta with parameters $\alpha > 0$ and $\beta > 0$. The plot of PDFs and CDFs of $Beta(\alpha,\, \beta)$ is provided in Figure 2.4. It is clear that if $\alpha = \beta = 1$, beta distribution coincides with $U(0,\, 1)$ distribution. The PDF of $Beta(\alpha,\, \beta)$ is unbounded at $x = 0$ and $x = 1$ if $\alpha < 1$ and $\beta < 1$, respectively. For $\alpha \geq 1$ and $\beta \geq 1$, the PDF of beta distribution is bounded. ||

Example 2.20 (RV which is neither discrete nor continuous)**.** Consider the RV X having CDF

$$F_X(x) = \begin{cases} 0 & \text{if} \quad x < -1 \\ x+1 & \text{if} \quad -1 \leq x < -1/2 \\ 1 & \text{if} \quad x \geq -1/2. \end{cases}$$

Note that the CDF is discontinuous only at 1/2. Now, $P(X = 1/2) = 1/2 \neq 1$. Hence, it is not a DRV. On the other hand it cannot be a CRV, as the CDF has discontinuity. Therefore, X is neither DRV nor CRV. Thus, there exists RV, which is not a DRV nor a CRV. Observe that $F_X(x) = \frac{1}{2}F_1(x) + \frac{1}{2}F_2(x)$ for all $x \in \mathbb{R}$, where $F_1(\cdot)$ and $F_2(\cdot)$ are distribution functions and given by

$$F_1(x) = \begin{cases} 0 & \text{if} \quad x < -1 \\ 2(x+1) & \text{if} \quad -1 \leq x < -1/2 \\ 1 & \text{if} \quad x \geq -1/2 \end{cases}$$

and

$$F_2(x) = \begin{cases} 0 & \text{if} \quad x < -1/2 \\ 1 & \text{if} \quad x \geq -1/2. \end{cases}$$

||

2.4 Expectation of Random Variable

Definition 2.7 (Expectation of DRV)**.** *Let X be a discrete RV with PMF $f_X(\cdot)$ and support S_X. The expectation or mean of X is defined by*

$$E(X) = \sum_{x \in S_X} x f_X(x) \quad \text{provided} \quad \sum_{x \in S_X} |x| f_X(x) < \infty.$$

If $\displaystyle\sum_{x \in S_X} |x| f_X(x) = \infty$ then we say that expectation does not exist.

Let us first try to understand the physical meaning of the expectation. This is an intuitive discussion and not mathematically flawless. However, it gives a good intuition of use of expectation. Let X be a discrete random variable that takes values $x_1,\, x_2,\, \ldots,\, x_k$ for some fixed value of k. Let we perform the corresponding random experiment N number of times. Let f_i^N denote frequency of x_i out of N trials. For example let a coin is tossed N times. There are two outcomes. Let X take value 1 if a head appear and 0 otherwise. Let f_1^N and

f_2^N, respectively, be the number of times heads and tails appear out of N trials. Then, the frequency of 1 is f_1^N and 0 is f_2^N.

The arithmetic average (or mean) of the observed outcomes is given by

$$Ave(N) = \frac{1}{N}\sum_{i=1}^{k} x_i f_i^N.$$

Let us see what will happen to $Ave(N)$ as $N \to \infty$. Note that

$$\lim_{N\to\infty} Ave(N) = \lim_{N\to\infty}\sum_{i=1}^{k} x_i \frac{f_i^N}{N} = \sum_{i=1}^{k} x_i \lim_{N\to\infty}\frac{f_i^N}{N} = \sum_{i=1}^{N} x_i p_i.$$

Consider the quantity $\lim_{N\to\infty}\frac{f_i^N}{N}$. This quantity can be interpreted as the proportion of times x_i observed out of N trials, when N is very large. Therefore, $\frac{f_i^N}{N}$ should, intuitively, converge to probability of $X = x_i$, which is denoted by p_i. This discussion shows that $E(X)$ can be interpreted as the long run mean (or average or weighted average) of the values that are assumed by a DRV.

The condition $\sum_{x\in S_X}|x|f_X(x) < \infty$ is imposed to ensure that the series $\sum_{x\in S_X} x f_X(x)$ converges absolutely. Note that a series $\sum_{n=1}^{\infty} x_n$ is called to converge absolutely if $\sum_{n=1}^{\infty}|x_n| < \infty$. Also, we know that if the series $\sum_{n=1}^{\infty}|x_n| < \infty$, then $\sum_{n=1}^{\infty} x_n$ converges. For expectation, $\sum_{x\in S_X} x f_X(x)$ absolutely converges if $\sum_{x\in S_X}|x f_X(x)| < \infty \implies \sum_{x\in S_X}|x|f_X(x) < \infty$, as $f_X(x) > 0$ for all $x \in S_X$.

If S_X is finite, $\sum_{x\in S_X}|x|f_X(x)$ is always finite. Therefore, $E(X)$ exists when S_X is finite. When S_X is countably infinite, we need to check if $\sum_{x\in S_X}|x|f_X(x)$ is finite or not.

Example 2.21. Let a fair die is rolled and let X denote the outcome of the roll. It is easy to see that X is a DRV with PMF

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1,\, 2,\, 3,\, 4,\, 5,\, 6 \\ 0 & \text{otherwise.} \end{cases}$$

Now, $\sum_{x\in S_X}|x|f_X(s) = \sum_{x=1}^{6}\frac{x}{6} < \infty$ as it is a finite sum. Hence, $E(X)$ exists and is given by

$$E(X) = \sum_{x=1}^{6}\frac{x}{6} = 3.5.$$

Note that the value of $E(X)$ does not belongs to the support of X. In other words, in this example $P\left(X = E(X)\right) = 0$. ||

Example 2.22. Let $X \sim Bin(n,\, p)$. In this case, the support is $S_X = \{0,\, 1,\, \ldots,\, n\}$, which is finite. Hence, $E(X)$ exists. The expectation can be calculated as follows.

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x} \\ &= \sum_{x=1}^{n} x\binom{n}{x}p^x(1-p)^{n-x} \\ &= \sum_{x=1}^{n} x \times \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \end{aligned}$$

$$
\begin{aligned}
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!\,((n-1)-(x-1))!} p^{x-1}(1-p)^{(n-1)-(x-1)} \\
&= np \times (p+1-p)^{n-1} \\
&= np.
\end{aligned}
$$

For $Bin(n,\, p)$ distribution, the success probability is p and we are performing n independent trials. If we interpret the probability as long term proportion, then out of n trials there will be approximately np successes. Hence, the value of the expectation is quite intuitive. ||

Example 2.23. Let $X \sim Geo(p)$. In this case, the support is $S_X = \{0,\, 1,\, \ldots\}$, which is a countably infinite set. Hence, first we need to check if the condition of the definition of expectation hold or not. To check it, we can proceed as follows:

$$
\sum_{x=0}^{\infty} |x| f_X(x) = \sum_{x=1}^{\infty} xp(1-p)^x = p \sum_{x=1}^{\infty} xq^x,
$$

where $q = 1-p$. Now, consider the partial sum

$$
S_n = \sum_{x=1}^{n} xq^x = \frac{q(1-q^n)}{p^2} - \frac{nq^{n+1}}{p} \to \frac{q}{p^2}
$$

as $n \to \infty$. Hence, $\sum_{x=0}^{\infty} |x| f_X(x) < \infty$ and the $E(X)$ exists. The expectation is given by

$$
E(X) = \sum_{x=0}^{\infty} x f_X(x) = \frac{q}{p}.
$$

||

Example 2.24. $X \sim Poi(\lambda)$. It can be shown that the expectation exists and $E(X) = \lambda$. Technique of showing it is similar to that of the above example. Therefore, I leave it for you to complete. ||

Example 2.25. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$
f(x) = \begin{cases} \frac{c}{x^2}, & x \in \mathbb{N}, \\ 0 & \text{otherwise}, \end{cases}
$$

where $c = \left(\sum_{n=1}^{\infty} \frac{1}{n^2} \right)^{-1}$. Clearly, $f(x) \geq 0$. The function $f(\cdot)$ is strictly greater than zero only on the set of natural numbers $\mathbb{N}$ and $\sum_{x=1}^{\infty} f(x) = 1$. Hence, $f(\cdot)$ is a PMF of a DRV, say X, with support $\mathbb{N}$, which is countably infinite. To compute $E(X)$, we first need to check if $\sum_{x=1}^{\infty} |x| f(x) < \infty$. Now,

$$
\sum_{x=0}^{\infty} |x| f(x) = c \sum_{n=1}^{\infty} \frac{1}{n},
$$

which is not finite. Hence, $E(X)$ does not exist in this case. ||

Definition 2.8 (Expectation of CRV)**.** *Let X be a CRV with PDF $f_X(\cdot)$. The expectation of X is defined by*

$$
E(X) = \int_{-\infty}^{\infty} x f_X(x) dx \quad \text{provided} \quad \int_{-\infty}^{\infty} |x| f_X(x) dx < \infty.
$$

Example 2.26. Let $X \sim U(a, b)$. First we need to check if $\int_a^b |x| \times \frac{1}{b-a} dx < \infty$.

For $a < b < 0$, $\int_a^b \frac{|x|}{b-a} dx = \int_a^b \frac{-x}{b-a} dx = -\frac{a+b}{2} < \infty$.

For $a < 0 \leq b$, $\int_a^b \frac{|x|}{b-a} dx = \int_a^0 \frac{-x}{b-a} dx + \int_0^b \frac{x}{b-a} dx = \frac{a^2+b^2}{2(b-a)} < \infty$.

For $0 \leq a < b$, $\int_a^b \frac{|x|}{b-a} dx = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2} < \infty$.

Hence, $E(X)$ exists and $E(X) = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}$. Loosely speaking, for a uniform distribution all points in (a, b) are equally likely, and hence, it is expected that the mean should be the middle point of the interval (a, b). ||

Example 2.27. Let $X \sim N(\mu, \sigma^2)$. Here, $E(X)$ exists for the following argument.

$$\begin{aligned}
\int_{-\infty}^{\infty} |x| f_X(x) dx &= \int_{-\infty}^{\infty} |x| \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |\mu + \sigma z| e^{-z^2/2} dz, \quad \text{taking } z = \frac{x-\mu}{\sigma} \\
&\leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(|\mu| + \sigma|z|\right) e^{-z^2/2} dz \\
&= \frac{\sqrt{2}}{\sqrt{\pi}} |\mu| \int_0^{\infty} e^{-z^2/2} dz + \frac{\sqrt{2}\sigma}{\sqrt{\pi}} \int_0^{\infty} z e^{-z^2/2} dz.
\end{aligned}$$

Now, for a $A > 0$, $\int_0^A z e^{-z^2/2} dz = 1 - e^{-A^2/2}$, which implies that $\int_0^{\infty} z e^{-z^2/2} dz < \infty$. Also, notice that

$$\int_0^{\infty} e^{-z^2/2} dz = \int_0^1 e^{-z^2/2} dz + \int_1^{\infty} e^{-z^2/2} dz \leq \int_0^1 e^{-z^2/2} dz + \int_1^{\infty} z e^{-z^2/2} dz < \infty.$$

Here, the first integration is a proper definite integration, where the integrand is bounded and continuous, and the range of integration is bounded. The second integration is finite by the argument used in the previous case. The expectation of X is given by

$$\begin{aligned}
\int_{-\infty}^{\infty} x f_X(x) dx &= \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z) e^{-z^2/2} dz, \quad \text{taking } z = \frac{x-\mu}{\sigma} \\
&= \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-z^2/2/} dz \\
&= \mu.
\end{aligned}$$

Here, the value of the integration in the first term is one, as the integrand is the PDF of a $N(0,\, 1)$ distribution. The value of the integration in the second term is zero, as the integrand is an odd function. ||

Example 2.28. Let X be a CRV having PDF $f_X(x) = \frac{1}{\pi(1+x^2)}$ for all $x \in \mathbb{R}$. The distribution of a CRV having this PDF is called Cauchy distribution. To check if $E(X)$ exist, we can proceed as follows.

$$\int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)}dx = \int_{-\infty}^{0} \frac{-x}{\pi(1+x^2)}dx + \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}dx.$$

Consider the second term. Notice that

$$\int_{0}^{A} \frac{x}{1+x^2}dx = \frac{1}{2}\ln(1+A^2) \to \infty \text{ as } A \to \infty.$$

Thus, $\int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)}dx$ is not finite, hence, $E(X)$ does not exist. Note that $\int_{-A}^{A} \frac{x}{1+x^2}dx = 0$ as the integrand is an odd function. Thus, $\int_{-\infty}^{\infty} \frac{x}{1+x^2}dx$ is conditionally integrable, but $\int_{-\infty}^{\infty} \frac{x}{1+x^2}dx$ is not integrable. ||

2.5 Transformation of Random Variable

Let $g : \mathbb{R} \to \mathbb{R}$ be a function and X be a RV. Then $Y = g(X)$ is also a RV. To see it, notice that a RV is a function from sample space to $\mathbb{R}$. Clearly, for $\omega \in \mathcal{S}$, $Y(\omega) = g(X(\omega)) \in \mathbb{R}$. In this section, our main aim is to find the distribution (CDF/PMF/PDF) of $Y = g(X)$, when the distribution of X is known. We will discuss mainly three techniques. In this section, we will also discuss about the expectation of a function of a RV.

2.5.1 Technique 1

In this technique, we will try to find the CDF of $Y = g(X)$ using the definition of CDF. That means that we will try to find $F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$ for all $y \in \mathbb{R}$. Note that CDF exists for all type of RVs. Therefore, this technique can be used for any type of RV. This technique is best understood by examples.

Example 2.29. Let the RV X has the following PMF:

$$f(x) = \begin{cases} \frac{1}{7} & \text{if } x = -2,\, -1,\, 0,\, 1 \\ \frac{3}{14} & \text{if } x = 2,\, 3 \\ 0 & \text{otherwise.} \end{cases}$$

Consider $Y = X^2$. Clearly, for $y < 0$, $F_Y(y) = P(X^2 \leq y) = P(X \in \emptyset) = 0$. For $y \geq 0$,

$$F_Y(y) = P\left(X^2 \leq y\right) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right).$$

Now, for $0 \leq y < 1$,

$$F_Y(y) = P(X = 0) = \frac{1}{7}.$$

For $1 \leq y < 4$,

$$F_Y(y) = P(X = 0 \text{ or } 1 \text{ or } -1) = \frac{3}{7}.$$

For $4 \leq y < 9$,

$$F_Y(y) = P(X = 0 \text{ or } 1 \text{ or } -1 \text{ or } 2 \text{ or } -2) = \frac{11}{14}.$$

For $y \geq 9$,

$$F_Y(y) = P(X = 0 \text{ or } 1 \text{ or } -1 \text{ or } 2 \text{ or } -2 \text{ or } 3) = 1.$$

Hence, the CDF of Y is

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{7} & \text{if } 0 \leq y < 1 \\ \frac{3}{7} & \text{if } 1 \leq y < 4 \\ \frac{11}{14} & \text{if } 4 \leq y < 9 \\ 1 & \text{if } y \geq 9. \end{cases}$$

In this case, Y is a DRV *(Why? Also, find the PMF of* Y*.).* ||

Example 2.30. Let the RV X has the following PDF:

$$f(x) = \begin{cases} \frac{|x|}{2} & \text{if } -1 < x < 1 \\ \frac{x}{3} & \text{if } 1 \leq x < 2 \\ 0 & \text{otherwise.} \end{cases}$$

Again consider the RV $Y = X^2$. For $y < 0$, $F_Y(y) = 0$. Like the previous example, for $y \geq 0$, $F_Y(y) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right)$. Now, for $0 \leq y < 1$,

$$F_Y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{|x|}{2} dx = \frac{y}{2}.$$

For $1 \leq y < 4$,

$$F_Y(y) = \int_{-1}^{1} \frac{|x|}{2} dx + \int_{1}^{\sqrt{y}} \frac{x}{3} dx = \frac{1}{6}\left(2 + y\right).$$

For $y \geq 4$,

$$F_Y(y) = \int_{-1}^{2} f(x) = 1.$$

It is clear that Y is a CRV *(Why? Also, find the PDF of* Y*.).* ||

Example 2.31. Let the RV X has the following PDF:

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

Suppose that we want to find the distribution of $Y = [X]$. Here, $[x]$ denotes the largest integer not exceeding x. First notice that $F_Y(y) = P\left(Y \leq y\right) = P\left([X] \leq y\right) = 0$ for all $y < 0$. For $0 \leq y < 1$,

$$F_Y(y) = P\left([X] \leq y\right) = P\left(X < 1\right) = \int_{-\infty}^{1} f(x)dx = \int_{0}^{1} e^{-x} dx = 1 - e^{-1}.$$

For $1 \leq y < 2$,

$$F_Y(y) = P\left([X] \leq y\right) = P\left(X < 2\right) = \int_{-\infty}^{2} f(x)dx = \int_{0}^{2} e^{-x}dx = 1 - e^{-2}.$$

In general, for $i \leq y < i+1$, where $i = 0,\, 1,\, 2, \ldots$,

$$F_Y(y) = P\left([X] \leq y\right) = P\left(X < i+1\right) = \int_{-\infty}^{i+1} f(x)dx = \int_{0}^{i+1} e^{-x}dx = 1 - e^{-(i+1)}.$$

Thus, the CDF of Y is given by

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - e^{-(i+1)} & \text{if } i \leq y < i+1, i = 0,\, 1,\, 2,\, \ldots. \end{cases}$$

Now, you can easily check that Y is a DRV by finding $P\left(Y = y\right)$ for all $y = 0,\, 1,\, 2,\, \ldots$ and then showing that $\sum_{y=0}^{\infty} P\left(Y = y\right) = 1$. Please complete. $||$

The basic idea here is to write the event $Y \leq y$ as $X \in A_y$ for appropriate set A_y. In the previous example, we have written $Y \leq y$ as $X \in (-\infty, i+1)$ for $y \in [i,\, i+1)$. Then using the distribution of X, one needs to find the probability of the event $X \in A_y$.

2.5.2 Technique 2

In Technique 2, we try to find PMF (if Y is DRV) or PDF (if Y is CRV) of Y directly without finding its' CDF. Obviously, first we need to understand whether Y is DRV or CRV. This technique is mainly based on two theorems. The first theorem consider the case when X is DRV. We will see that if X is DRV, then Y is also a DRV. The second theorem addresses the case when X is CRV. We will see that under some conditions, Y is a CRV if X is CRV. With examples, we will illustrate that if the conditions do not hold, then Y can be DRV as well as CRV. Hence, those conditions are important. Let us start with an example.

Example 2.32. Let the RV X has the following PMF:

$$f(x) = \begin{cases} \frac{1}{7} & \text{if } x = -2,\, -1,\, 0,\, 1 \\ \frac{3}{14} & \text{if } x = 2,\, 3 \\ 0 & \text{otherwise.} \end{cases}$$

Consider $Y = X^2$ and suppose that we want to find PMF or PDF, whatever applicable, of Y. Note that the support of X is $S_X = \{-2,\, -1,\, 0,\, 1,\, 2,\, 3\}$. Intuition says that Y should takes value from the set $D = \{0,\, 1,\, 4,\, 9\}$ with positive probabilities. Based on this intuition, we will try to find $P\left(Y = y\right)$ for all $y \in D$ and then check if $\sum_{y \in D} P\left(Y = y\right)$ equal one or not.

$$P\left(Y = 0\right) = P\left(X = 0\right) = \frac{1}{7}.$$

$$P\left(Y = 1\right) = P\left(X = 1 \text{ or } -1\right) = \frac{2}{7}.$$

$$P\left(Y = 4\right) = P\left(X = 2 \text{ or } -2\right) = \frac{5}{14}.$$

$$P(Y=9) = P(X=3 \text{ or } -3) = \frac{3}{14}.$$

Note that again to compute $P(Y=y)$, we first find the inverse image of $Y=y$ as $X \in A_y$ and then used the distribution of X. Thus, $A_y = \{x \in \mathbb{R} : x^2 = y\}$. In the last case, $P(Y=9)$, suggests that even we do not need to consider all the elements x such that $x^2 = 9$. We need to only consider those x, which are in S_X and $x^2 = y$. Thus, we can take $A_y = \{x \in S_X : x^2 = y\}$. It is clear that $\sum_{y \in D} P(Y=y) = 1$. Hence, Y is a DRV with support D and PMF

$$f(y) = \begin{cases} \frac{1}{7} & \text{if } y = 0 \\ \frac{2}{7} & \text{if } y = 1 \\ \frac{5}{14} & \text{if } y = 4 \\ \frac{4}{14} & \text{if } y = 9 \\ 0 & \text{otherwise.} \end{cases}$$

||

Theorem 2.7. *Let X be a DRV with PMF $f_X(\cdot)$ and support S_X. Let $g : \mathbb{R} \to \mathbb{R}$ and $Y = g(X)$. Then Y is a DRV with support $S_Y = \{g(x) : x \in S_X\}$ and PMF*

$$f_Y(y) = \begin{cases} \sum_{x \in A_y} f_X(x) & \text{if } y \in S_Y \\ 0 & \text{otherwise,} \end{cases} \tag{2.10}$$

where $A_y = \{x \in S_X : g(x) = y\}$.

Proof: For $y \in g(S_X)$, $P(Y=y) = P(X \in A_y) = \sum_{x \in A_y} f_X(x)$. Also, $g(S_X)$ is atmost countable as S_X is at most countable. Now, we will try to show that $\sum_{y \in g(S_X)} P(Y=y) = 1$ or equivalently $\sum_{y \in g(S_X)} \sum_{x \in A_y} f_X(x) = 1$. Notice that

$$\bigcup_{y \in g(S_X)} A_y = S_X.$$

To see it first assume that $x \in S_X$ which implies that $y = g(x) \in g(S_X)$. Hence, for some $y \in g(S_X)$, $x \in A_y$. On the other hand, if $x \in \bigcup_{y \in g(S_X)} A_y$, then for some $y \in g(S_X)$, $x \in A_y \subset S_X$. Hence, $x \in S_X$.

Next notice that

$$A_{y_1} \bigcap A_{y_2} = \emptyset \quad \text{for } y_1 \neq y_2 \in g(S_X).$$

We can prove this claim by contradiction. Suppose that $A_{y_1} \bigcap A_{y_2} \neq \emptyset$. That means there exists at least one $x \in A_{y_1}$ and $x \in A_{y_2}$. Hence, $y_1 = g(x) = y_2$, which is a contradiction to the fact that $y_1 \neq y_2$. Thus,

$$\sum_{y \in g(S_X)} \sum_{x \in A_y} f_X(x) = \sum_{x \in S_X} f_X(x) = 1.$$

Hence, Y is a DRV with support $S_Y = g(S_X)$ and PMF as given in (2.10) □

Example 2.33. Let $X \sim Bin(n,\, p)$. Suppose that we are interested to find the distribution of $Y = n - X$. As X is a DRV, using the above theorem, Y is also DRV. Here, $S_X = \{0,\, 1,\, \ldots,\, n\} = S_Y$. For any $y \in S_Y$, $A_y = \{n - y\}$. Hence, the PMF of Y is

$$\begin{aligned} f_Y(y) &= \begin{cases} f_X(n-y) & \text{if } y = 0,\, 1,\, \ldots,\, n \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \binom{n}{n-y} p^{n-y}(1-p)^{n-n+y} & \text{if } y = 0,\, 1,\, \ldots,\, n \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \binom{n}{y}(1-p)^y p^{n-y} & \text{if } y = 0,\, 1,\, \ldots,\, n \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Hence, $Y \sim Bin(n,\, 1-p)$. Note that $Y = n - X$ is the number of failures out of n trials. Therefore, this result is well justified. ||

Theorem 2.8. *Let X be a CRV with PDF $f_X(\cdot)$ and support S_X, which is an interval. Let $g : S_X \to \mathbb{R}$ be a differentiable function and either $g'(x) < 0$ for all $x \in S_X$ or $g'(x) > 0$ for all $x \in S_X$. Then the RV $Y = g(X)$ is a CRV with PDF*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left|\frac{d}{dy} g^{-1}(y)\right| & \textit{for } y \in g(S_X) \\ 0 & \textit{otherwise.} \end{cases}$$

Proof: The proof of the theorem can be done using the results of transformation of variable technique in integration. However, it is skipped here. □

Example 2.34. Let $X \sim U(0,\, 1)$. Suppose that $g(x) = -\ln x$ for $x \in (0,\, 1)$. Also, the support of X is $S_X = (0,\, 1)$, which is an interval. Clearly, $g'(x) < 0$ for all $x \in (0,\, 1)$. The inverse of $g(\cdot)$ is $g^{-1}(y) = e^{-y}$ for all $y \in g(S_X) = (0,\, \infty)$. Hence, $Y = -\ln X$ is a CRV with PDF

$$\begin{aligned} f_Y(y) &= \begin{cases} f_X\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right| & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} e^{-y} & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Therefore, $Y = -\ln X \sim Exp(1)$. ||

Example 2.35. Let $X \sim Exp(1)$. Suppose we are interested to find the distribution of $Y = X^2$. Here, $g(x) = x^2$ for $x \in S_X = (0,\, \infty)$. Also, $g'(x) = 2x > 0$ for all $x > 0$. Hence, $Y = X^2$ is a CRV. Note that $g^{-1}(y) = \sqrt{y}$. Thus, the PDF of Y is

$$f_Y(y) = \begin{cases} e^{-\sqrt{y}} \times \frac{1}{2\sqrt{y}} & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that in this case the function $g(x) = x^2$ defined on $\mathbb{R}$ is not strictly monotone. However, we need to check only on the support of X and $g(\cdot)$ is strictly monotone on $(0,\, \infty)$. ||

Example 2.36. Let $X \sim N(0, 1)$. Suppose that we want to find the distribution of $Y = X^2$. Note that the support of X is $\mathbb{R}$ and $g'(x) = 2x$ does not take only positive or negative values on $\mathbb{R}$. Hence, we cannot use Theorem 2.8. However, we can use technique 1 to obtain the CDF of Y and then check the type of the RV Y. The CDF of Y is given by

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 2\Phi\left(\sqrt{y}\right) - 1 & \text{if } y \geq 0. \end{cases}$$

It is easy to see that $F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt$, where

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{y}}\phi\left(\sqrt{y}\right) & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, Y is a CRV. This example shows that even if some of the conditions of the Theorem 2.8 do not hold true, the RV Y could be CRV. On the other hand, consider the Example 2.31, where $g(x) = [x]$. This function also does not satisfy the strictly monotone condition of the Theorem 2.8 and we have seen that Y is a DRV. Thus, the conditions in the Theorem 2.8 are important and they are sufficient conditions, but not necessary. ||

2.5.3 Expectation of Function of RV

In this subsection, we will consider the expectation of a function of a RV. Let X be a RV and $g : S_X \to \mathbb{R}$ be a function. Then we have seen that $Y = g(X)$ is a RV. Naturally, we may want talk about the expectation of Y, if it exist. Note that one of the way to check if $E(Y)$ exists and then to compute it is to find the PMF or PDF of Y and then use the definition of expectation as we have done earlier. Let us consider an example.

Example 2.37. Let X be a DRV with PMF

$$f_X(x) = \begin{cases} \frac{1}{7} & \text{if } x = -2,\ -1,\ 0,\ 1 \\ \frac{3}{14} & \text{if } x = 2,\ 3 \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = X^2$. In Example 2.32, we have seen that the PMF of Y is

$$f(y) = \begin{cases} \frac{1}{7} & \text{if } y = 0 \\ \frac{2}{7} & \text{if } y = 1 \\ \frac{5}{14} & \text{if } y = 4 \\ \frac{4}{14} & \text{if } y = 9 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, we can compute the expectation of Y based on this PMF. Note that in this case S_Y is finite, and hence, $E(Y)$ exists. The expectation of Y is given by

$$E(Y) = 0 \times \frac{1}{7} + 1 \times \frac{2}{7} + 4 \times \frac{5}{14} + 9 \times \frac{4}{14} = \frac{30}{7}.$$

However, we can compute the expectation without computing the PMF of Y. Notice that

$$\begin{aligned}
E(Y) &= 0\times\frac{1}{7}+1\times\frac{2}{7}+4\times\frac{5}{14}+9\times\frac{4}{14}\\
&= 0^2\times P(X=0)+1^2\times P(X=1)+(-1)^2\times P(X=-1)\\
&\qquad +2^2\times P(X=2)+(-2)^2\times P(X=-2)+3^2\times P(X=3)\\
&= \sum_{x\in S_X} x^2 P(X=x).
\end{aligned}$$

The benefit of this is that we do not need to compute the PMF of Y. We can use the PMF of X to compute the $E(Y)=E(X^2)$. ||

Theorem 2.9. *Let X be a DRV with PMF $f_X(\cdot)$ and support S_X. Let $g:\mathbb{R}\to\mathbb{R}$. Then*

$$E\left[g(X)\right]=\sum_{x\in S_X} g(x)f_X(x)\quad \textit{provided} \sum_{x\in S_X}|g(x)|f_X(x)<\infty.$$

Proof: Using the Theorem 2.7,

$$E(Y)=\sum_{y\in S_Y} y\sum_{x\in A_y} f_X(x)=\sum_{y\in S_Y}\sum_{x\in A_y} g(x)f_X(x)=\sum_{x\in S_X} g(x)f_X(x).$$

□

Theorem 2.10. *Let X be a CRV with PDF $f_X(\cdot)$. Let $g:\mathbb{R}\to\mathbb{R}$. Then*

$$E\left[g(X)\right]=\int_{-\infty}^{\infty} g(x)f_X(x)dx\quad \textit{provided} \int_{-\infty}^{\infty}|g(x)|f_X(x)dx<\infty.$$

Proof: The proof of this theorem is out of the scope of this course. □

As we have pointed out that $Y=g(X)$ may be a DRV or a CRV if X is a CRV. However, the previous theorem tells us that whatever the type of the RV Y is, we can use the theorem if X is a CRV.

Example 2.38 (Continuation of Example 2.32)**.** Let $X\sim U(0,1)$. Then the expectation of $Y=-\ln X$ exists if $\int_0^1|-\ln x|dx=-\int_0^1\ln x dx<\infty$. Now,

$$\int_\varepsilon^1 \ln x dx=-1-\varepsilon\ln\varepsilon+\varepsilon\to -1\quad\text{as } \varepsilon\downarrow 0.$$

Hence, $E(Y)$ exists and is given by

$$E(Y)=-\int_0^1 \ln x dx=1.$$

||

Example 2.39 (Continuation of Example 2.31)**.** Let $X\sim Exp(1)$. The expectation of $Y=[X]$ exists if

$$\int_0^\infty |[x]|\,e^{-x}dx=\sum_{i=0}^\infty\int_i^{i+1}|[x]|\,e^{-x}dx=\sum_{i=0}^\infty i\left(e^{-i}-e^{-(i+1)}\right)<\infty.$$

Now, using the technique that is used in Example 2.23, we can show that $\sum_{i=0}^{\infty} ie^{-i} < \infty$, which implies that $E(Y)$ exists. The expectation of Y is given by

$$\int_0^{\infty} [x]e^{-x}dx = \sum_{i=0}^{\infty} \int_i^{i+1} [x]e^{-x}dx = \sum_{i=0}^{\infty} i\left(e^{-i} - e^{-(i+1)}\right) = \frac{e^{-1} - e^{-2}}{\left(1 - e^{-1}\right)^2}.$$

||

Theorem 2.11. *Let X be a RV (either DRV or CRV).*

1. *Let $A \subset \mathbb{R}$. Then $E(I_A(X)) = P(X \in A)$, where I_A denotes the indicator function of the event A.*

2. *Let $h_1(x) \leq h_2(x)$ for all $x \in \mathbb{R}$ be two real valued functions. Then $E\left[h_1(X)\right] \leq E\left[h_2(X)\right]$, provided all the expectations exist.*

3. *Let $a < b$ be two fixed real numbers such that $S_X \subset [a,\, b]$. Then $a \leq E(X) \leq b$, provided the expectation exists.*

4. *$E(a + bX) = a + bE(X)$, where a and b are two fixed real numbers.*

5. *Let $h_1(\cdot),\, \ldots,\, h_p(\cdot)$ be real valued functions of real numbers such that $E(h_i(X))$ exists for all $i = 1,\, 2, \ldots,\, p$, then*

$$E\left(\sum_{i=1}^{p} h_i(X)\right) = \sum_{i=1}^{p} E\left(h_i(X)\right).$$

Proof: The proof this theorem is straight forward and therefore, left as an exercise. □

Definition 2.9 (Raw Moment)**.** *For $r = 1,\, 2,\, \ldots,\, \mu_r = E(X^r)$ is called rth raw moment of X, if the expectation exists.*

Definition 2.10 (Central Moment)**.** *$\mu_r' = E\left[(X - E(X))^r\right]$ is called rth central moment of X, if the expectations exist.*

Definition 2.11 (Variance)**.** *$\mu_2' = E\left[(X - E(X))^2\right]$ is called variance of X when it exists and is denoted by $Var(X)$. Note that $Var(X) = E(X^2) - (E(X))^2$.*

Theorem 2.12. *$E\left(X - E(X)\right)^2 \leq E\left(X - a\right)^2$ for all $a \in \mathbb{R}$.*

Proof: Let us denote $\mu = E(X)$. Then

$$\begin{aligned} E\left(X - a\right)^2 &= E\left(X - \mu + \mu - a\right)^2 \\ &= E\left(X - \mu\right)^2 + E\left(\mu - a\right)^2 + 2E\left[(X - \mu)(\mu - a)\right] \\ &= E\left(X - \mu\right)^2 + \left(\mu - a\right)^2 + 2(\mu - a)E\left(X - \mu\right) \\ &= E\left(X - \mu\right)^2 + \left(\mu - a\right)^2 \\ &\geq E\left(X - \mu\right)^2. \end{aligned}$$

The third equality is due to the fact that $\mu - a$ is a constant. The fourth equality holds true as $E\left(X - \mu\right) = E(X) - \mu = 0$. Finally, the last inequality is due to the fact that $(\mu - a)^2 \geq 0$. Note that $E\left(X - a\right)^2 = E\left(X - \mu\right)^2$ if and only if $\left(\mu - a\right)^2 = 0$, which implies and is implied by $a = \mu$. □

Note that $E\left(X-a\right)^2$ can be regarded as the average error if we use a instead of X. Thus, if we do not know the value of X and need to provide a guess for the same, then it is safer to use $E(X)$ as a guess than any other value, as the average error is minimum for $E\left(X\right)$. Therefore, $E(X)$ is the "best estimate" of X.

Definition 2.12 (Moment Generating Function)**.** *The moment generating function (MGF) of a RV X is defined by*

$$M_X(t) = E\left(e^{tX}\right)$$

provided there exists a real number $a > 0$ such that the expectation exists for all $t \in (-a,\, a)$.

Example 2.40. Let $X \sim Bin(n,\, p)$. Then

$$E\left(e^{tX}\right) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} = \sum_{x=0}^{n}\binom{n}{x}\left(pe^t\right)^x(1-p)^{n-x} = (1-p+pe^t)^n$$

for all $t \in \mathbb{R}$. As the sum is a finite sum, the sum converges for all $t \in \mathbb{R}$. As $E\left(e^{tX}\right)$ exists for all $t \in \mathbb{R}$, we can take any value of $a > 0$. However, we write that MGF exists for all $t \in \mathbb{R}$ in such situations. Therefore, the MGF of X is given by

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

for all $t \in \mathbb{R}$. ||

Example 2.41. Let $X \sim Exp(\lambda)$. Then

$$E\left(e^{tX}\right) = \lambda\int_0^\infty e^{tx}e^{-\lambda x}dx = \lambda\int_0^\infty e^{-(\lambda-t)x}dx.$$

This integration converges if and only if $t < \lambda$. Hence,

$$E\left(e^{tX}\right) = \left(1-\frac{t}{\lambda}\right)^{-1}$$

for all $t < \lambda$. Clearly, we can take $a = \lambda > 0$ and $E\left(e^{tX}\right)$ exists for all $t \in (-\lambda,\, \lambda)$. Thus, MGF of X exists and is given by

$$M_X(t) = \left(1-\frac{t}{\lambda}\right)^{-1}$$

for all $t < \lambda$. Notice that the range of t, for which $E\left(e^{tX}\right)$ exists, is important and need to specify unambiguously. ||

Example 2.42. Let $X \sim N(\mu,\, \sigma^2)$. Then

$$\begin{aligned}
E\left(e^{tX}\right) &= \int_{-\infty}^{\infty} e^{tx}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\
&= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2\sigma^2}\left(x^2-2\mu x+\mu^2-2\sigma^2tx\right)\right]dx \\
&= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2\sigma^2}\left(x^2-2\left(\mu+t\sigma^2\right)x+\left(\mu+t\sigma^2\right)^2-\left(\mu+t\sigma^2\right)^2+\mu^2\right)\right]dx
\end{aligned}$$

$$= \exp\left[-\frac{1}{2\sigma^2}\left(\mu^2 - \mu^2 - t^2\sigma^4 - 2\mu\sigma^2 t\right)\right]\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}\left(x-\mu-t\sigma^2\right)^2\right]dx$$
$$= \exp\left[\mu t + \frac{1}{2}t^2\sigma^2\right]$$

for all $t \in \mathbb{R}$. Here, the last equality follows from the fact that the integrand (in the last but one line) is the PDF of a $N(\mu + t\sigma^2,\, \sigma^2)$ distribution. Therefore, the MGF of X is

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

for all $t \in \mathbb{R}$. ||

Theorem 2.13. *If the MGF $M_X(t)$ exist for $t \in (-a,\, a)$ for some $a > 0$, the derivatives of all order exist at $t = 0$ and*

$$E\left(X^k\right) = \left.\frac{d^k}{dt^k}M_X(t)\right|_{t=0}$$

for all positive integer k.

Proof: Out of the scope of this course. □

Though the proof of the previous theorem is out of scope, it is quite important. Note that this theorem tells us that if MGF exists then all raw moment, and hence, all central moments exists. Also, we can compute raw moments using the MGF, and hence, the name MGF. Moreover, we need to take rth derivative of MGF at zero to compute rth raw moments. Also, for the main theorem of the next subsection, we need the condition that two MGFs have to be equal in a neighborhood of zero. These are the reasons of keeping the condition that the MGF exists if $E\left(e^{tX}\right)$ exists in a neighborhood of zero in the definition of MGF. Note that if $E\left(e^{tX}\right)$ exists only on an interval not including zero, then $E\left(e^{tX}\right)$ is of no use.

2.5.4 Technique 3

Definition 2.13 (Same in Distribution)**.** *Two RVs X and Y are said to be same in distribution (denoted by $X \overset{d}{=} Y$) if $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.*

Theorem 2.14. *Let X and Y be two RVs having MGFs $M_X(\cdot)$ and $M_Y(\cdot)$, respectively. Suppose that there exists a positive real number a such that $M_X(t) = M_Y(t)$ for all $t \in (-a,\, a)$. Then X and Y are same in distribution.*

Proof: The proof of this theorem is out of scope of this course. □

In general, for some function $E\left(f(X)\right) = E\left(f(Y)\right)$ does not imply that $X \overset{d}{=} Y$. However, MGF is very special in this respect. Note that this theorem can be use to find the distribution of a function of a RV as illustrated in the following example.

Example 2.43. Let $X \sim N(\mu,\, \sigma^2)$. Suppose we are interested to find the distribution of $Y = a + bX$, which is a linear combination of X. Assume that $b \neq 0$. Otherwise $Y = a$ with probability one. First, let us try to find the MGF of Y. Note that

$$E\left(e^{tY}\right) = E\left(e^{t(a+bX)}\right) = e^{ta}E\left(e^{tbX}\right) = e^{ta}M_X(tb).$$

Now, from the Example 2.42, $M_X(t)$ exists for all $t \in \mathbb{R}$. Hence,

$$E(e^{tY}) = e^{ta}e^{\mu bt+\frac{1}{2}b^2t^2\sigma^2} = e^{(a+b\mu)t+\frac{1}{2}(b\sigma)^2t^2}$$

for all $t \in \mathbb{R}$. Suppose that $Z \sim N(a + b\mu,\, b^2\sigma^2)$. Then the MGF of Z is

$$M_Z(t) = e^{(a+b\mu)t+\frac{1}{2}b^2\sigma^2t^2}$$

for all $t \in \mathbb{R}$. Thus, the MGFs of Y and Z are same for all $t \in \mathbb{R}$. Thus, $Y \overset{d}{=} Z \sim N(a + b\mu,\, b^2\sigma^2)$. Note that to use the technique 3, we need to identify the MGF of Y. ||

2.6 Moment Inequality

In this section we will discuss some inequalities involving probability and moment of a random variables. These inequalities give upper bound of probability of the events of the form $|X| \ge c$ in terms of moment of X. Obviously, if the upper bound is greater than or equal to one, then these inequality do not provide any extra information. Therefore, these inequalities are meaningful if the upper bond turn out to be less than one.

Theorem 2.15. *Let X be a RV and $g : [0,\, \infty) \to [0,\, \infty)$ be a non-decreasing function such that $E\left(g(|X|)\right)$ is finite. Then for any $c > 0$ with $g(c) > 0$,*

$$P(|X| \ge c) \le \frac{E\left(g(|X|)\right)}{g(c)}.$$

Proof: Notice that for all $x \in \mathbb{R}$,

$$g(c)I_{(-\infty,\, c]\cup[c,\, \infty)}(x) \le g(|x|), \tag{2.11}$$

where

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

To see it, first let $|x| \ge c$. Then the left hand side is $g(c)$ and the right hand side is $g(|x|)$. As $g(\cdot)$ is a non-decreasing function, (2.11) holds true for $|x| \ge c$. For $-c < x < c$, the left hand side is zero and right hand side is $g(|x|)$. As the range of $g(\cdot)$ in non-negative part of real line, (2.11) holds true for $-c < x < c$. Thus, (2.11) holds true for all $x \in \mathbb{R}$. Now, using Theorem 2.11,

$$E\left[g(c)I_{(-\infty,\, c]\cup[c,\, \infty)}(X)\right] \le E\left(g\left(|X|\right)\right) \implies g(c)P\left(|X| \ge c\right) \le E\left(g\left(|X|\right)\right).$$

Now, as $g(c) > 0$, proof of the theorem completes. □

Corollary 2.1 (Markov Inequality). *Let X be a RV with $E\left(|X|^r\right) < \infty$ for some $r > 0$. Then for any $c > 0$,*

$$P(|X| \ge c) \le \frac{E\left(|X|^r\right)}{c^r}.$$

Proof: Take $g(x) = x^r$ in the previous theorem. □

Corollary 2.2 (Chebyshev Inequality). *Let X be a RV with $E(X^2) < \infty$. Let us denote $\mu = E(X)$ and $\sigma^2 = Var(X)$. Then for any $k > 0$,*

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

Proof: Let $Y = X - \mu$. Taking $r = 2$ and $c = k$ in the previous corollary, we get

$$P\left(|Y| \geq k\right) \leq \frac{E\left(|Y|^2\right)}{k^2} \implies P\left(|X - \mu| \geq k\right) \leq \frac{E\left((X-\mu)^2\right)}{k^2} = \frac{\sigma^2}{k^2}.$$

□

Example 2.44. Let X be a DRV with PMF

$$f_X(x) = \begin{cases} \frac{1}{8} & \text{if } x = -1,\, 1 \\ \frac{3}{4} & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then $E(X) = 0$ and $E(X^2) = 1/4$, which implies that $Var(X) = 1/4$. Now, using Chebyshev inequality and taking $k = 1$, $P\left(|X| \geq 1\right) \leq \frac{1}{4}$. On the other hand, using PMF of X, $P\left(|X| \geq 1\right) = \frac{1}{4}$. This example shows that in general we do not have a better upper bound of $P\left(|X| \geq k\right)$ than the upper bound provided by Chebyshev inequality. Thus, the Chebyshev inequality is tight. Of course, if we have extra information on the RV X, then it may be possible to have a better upper bound of $P\left(|X| \geq k\right)$ compared to that is given by Chebyshev inequality. ||

2.A Gamma Integral

Lemma 2.1. *The improper integral*

$$\int_0^\infty t^{n-1} e^{-t} dt$$

converges absolutely for all $n \in \mathbb{N}$.

Proof: As $\lim\limits_{t \to \infty} t^{n-1} e^{-t/2} = 0$ for all $n \in \mathbb{N}$, there exists $t_0 > 0$ such that $t^{n-1} \leq e^{t/2}$ for all $t > t_0$. Now,

$$\int_0^\infty t^{n-1} e^{-t} dt = \int_0^{t_0} t^{n-1} e^{-t} dt + \int_{t_0}^\infty t^{n-1} e^{-t} dt,$$

where the first integral is a proper integral and the second is an improper integral. The second integral converges as

$$\int_{t_0}^\infty t^{n-1} e^{-t} dt \leq \int_{t_0}^\infty e^{-t/2} dt = \lim_{A \to \infty} \int_{t_0}^A e^{-t/2} dt = 2e^{-\frac{t_0}{2}} < \infty.$$

□

Lemma 2.2. *The improper integral*

$$\int_0^\infty t^{\alpha-1} e^{-t} dt$$

converges absolutely for all $\alpha \geq 1$.

Proof: For $t > 1$, $t^{\alpha-1}e^{-t} \leq t^{[\alpha]}e^{-t}$, where $[x]$ denotes the largest integer not exceeding $x \geq 0$. Now,

$$\int_0^\infty t^{\alpha-1}e^{-t}dt = \int_0^1 t^{\alpha-1}e^{-t}dt + \int_1^\infty t^{\alpha-1}e^{-t}dt,$$

where the first integral is a proper integral and the second is an improper integral. The second integral converges as

$$\int_1^\infty t^{\alpha-1}e^{-t}dt \leq \int_1^\infty t^{[\alpha]}e^{-t/2}dt < \int_0^\infty t^{[\alpha]}e^{-t}dt,$$

which converges by the Lemma 2.1. □

Lemma 2.3. *The improper integral*

$$\int_0^\infty t^{\alpha-1}e^{-t}dt$$

converges absolutely for all $0 < \alpha < 1$.

Proof: Notice that $t^{\alpha-1}e^{-t} \leq e^{-t}$ for $t > 1$. Now,

$$\int_0^\infty t^{\alpha-1}e^{-t}dt = \int_0^1 t^{\alpha-1}e^{-t}dt + \int_1^\infty t^{\alpha-1}e^{-t}dt,$$

where both the integrals on the right hand side are improper. Consider the first integral. Notice that $e^{1+t} \geq 1 \implies e^{-t} \leq e$ for $t \geq 0$. Hence,

$$\int_0^1 t^{\alpha-1}e^{-t}dt \leq e\int_0^1 t^{\alpha-1} < \infty.$$

Now, consider the second integral.

$$\int_1^\infty t^{\alpha-1}e^{-t}dt \leq \int_1^\infty e^{-t}dt < \infty.$$

Thus, both the integral converges, which proves the Lemma. □

Theorem 2.16. *The improper integral*

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$$

converges absolutely for all $\alpha > 0$.

Proof: Combining Lemmas 2.1, 2.2, and 2.3, the proof of the theorem is immediate. □

Theorem 2.17. *The functional equation* $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ *holds for* $\alpha > 0$.

Proof:

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty t^{\alpha}e^{-t}dt \\ &= \left[t^\alpha \int e^{-t}dt + \int \alpha t^{\alpha-1}e^{-t}dt\right]_0^\infty \\ &= \alpha\int_0^\infty t^{\alpha-1}e^{-t}dt \\ &= \alpha\Gamma(\alpha).\end{aligned}$$

□

Theorem 2.18. $\Gamma(n+1) = n!$ *for* $n = 1, 2, \ldots$.

Proof: It is very easy to see (using integration by parts) that the theorem holds for $n = 1$. Now, using the previous theorem, proof of current theorem is immediate. □

2.B Beta Integral

Theorem 2.19. *For* $\alpha > 0$ *and* $\beta > 0$*, the integral*

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

converges.

Proof: Note that for $\alpha \geq 1$ and $\beta \geq 1$, the integral is a proper integral. The integral is improper if $0 < \alpha < 1$ and/or $0 < \beta < 1$. If $\alpha < 1$, $f(x) \to \infty$ as $x \downarrow 0$, where $f(\cdot)$ is the integrand. Similarly, $f(x) \to \infty$ as $x \uparrow 1$ if $\beta < 1$. First notice that

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \int_0^{1/2} x^{\alpha-1}(1-x)^{\beta-1}dx + \int_{1/2}^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

The first and second integrals on the right hand side are improper if $\alpha < 1$ and $\beta < 1$, respectively. Let us first prove that the first integral converges. Note that for $0 < x < 1/2$, $(1-x)^{\beta-1} \leq A(\beta)$, where

$$A(\beta) = \begin{cases} 1 & \text{if } \beta \geq 1 \\ \left(\frac{1}{2}\right)^{\beta-1} & \text{if } 0 < \beta < 1. \end{cases}$$

Thus,

$$\int_0^{1/2} x^{\alpha-1}(1-x)^{\beta-1}dx \leq A(\beta) \int_0^{1/2} x^{\alpha-1}dx < \infty.$$

Similarly, we can prove that the second integral converges by deducing an upper bound of $x^{\alpha-1}$ for $1/2 < x < 1$. □

Theorem 2.20. *For* $\alpha > 0$ *and* $\beta > 0$,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Proof:

$$\begin{aligned}
\Gamma(\alpha+\beta)B(\alpha, \beta) &= \left(\int_0^\infty x^{\alpha+\beta-1}e^{-x}dx\right)\left(\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy\right) \\
&= \int_0^\infty \int_0^1 x^{\alpha+\beta-1}e^{-x}y^{\alpha-1}(1-y)^{\beta-1}dydx \\
&= \int_0^\infty \int_0^\infty z_1^{\alpha-1}z_2^{\beta-1}e^{-(z_1+z_2)}dz_1dz_2, \quad \text{taking } z_1 = xy \text{ and } z_2 = (1-y)x \\
&= \Gamma(\alpha)\Gamma(\beta).
\end{aligned}$$

□

Corollary 2.3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Proof: Using the last theorem,

$$\left(\Gamma\left(\frac{1}{2}\right)\right)^2 = B\left(\frac{1}{2}, \frac{1}{2}\right) = \int_0^1 \frac{1}{\sqrt{x(1-x)}} dx = \pi.$$

As $\Gamma(1/2)$ is integration of a positive function, $\Gamma(1/2) > 0$. Hence, $\Gamma(1/2) = \sqrt{\pi}$. □

Chapter 3

Jointly Distributed Random Variables

In the previous chapter, we studied RV and associated concepts. One of the main uses of the random variable is to model numerical characteristic of a natural phenomena. For example, we may assume that the RV X denote the income of a household. Now, assume that Y denotes the spending of the household. Then $Z = X - Y$ is a RV and it is the savings of the household. Clearly, if we know the values of any of two RVs among X, Y, and Z, the value of other one is known. In such situation, we may want to study the relationship between X and Z. Clearly, the concepts of the previous chapter are not sufficient. Using the tools of the previous chapter, we can study X and Z separately, but not jointly. For example we cannot answer the question: Is savings increases with income? To answer it, we need to know the probability of joint occurrence of events $X \leq x$ and $Z \leq z$ for different values of x and z. In this chapter, we will study the concepts relating to the joint occurrence of multiple RVs. There are plenty of examples, where we need to consider multiple RVs. These examples include height and weight of a person, pollution level and blood pressure, lifetime of a product and its cause of failure, etc.

3.1 Random Vector

Definition 3.1 (Random Vector)**.** *A function $\boldsymbol{X} : \mathcal{S} \to \mathbb{R}^n$ is called a random vector.*

Clearly, random vector is a generalization of RV. Note that as $\boldsymbol{X}$ is a function from $\mathcal{S}$ to $\mathbb{R}^n$, $\boldsymbol{X}(\omega)$ can be written as $(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$, where $X_i : \mathcal{S} \to \mathbb{R}$ for all $i = 1, 2, \ldots, n$. Thus, X_i is a RV for all $i = 1, 2, \ldots, n$. Therefore, each component of a random vector is a RV and we will write $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$.

Definition 3.2 (Joint CDF)**.** *For any random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, the joint cumulative distribution function (JCDF) is defined by*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P\left(X_1 \leq x_1, \ldots, X_n \leq x_n\right),$$

for all $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Here

$$\{X_1 \leq x_1, \ldots, X_n \leq x_n\} = \{X_1 \leq x_1\} \cap \{X_2 \leq x_2\} \cap \ldots \cap \{X_n \leq x_n\}.$$

Now, onward, most of the definitions, theorems, results will be presented for $n = 2$. Thus, we will use $\boldsymbol{X} = (X_1, X_2)$ or $\boldsymbol{X} = (X, Y)$. This is for simplicity of the expressions. However, most of the definitions, theorems, results can be extended for any general value of n. For $n = 2$, the JCDF at the point (x, y) is the probability that the random vector (X, Y) belongs to the shaded region of the Figure 3.1.

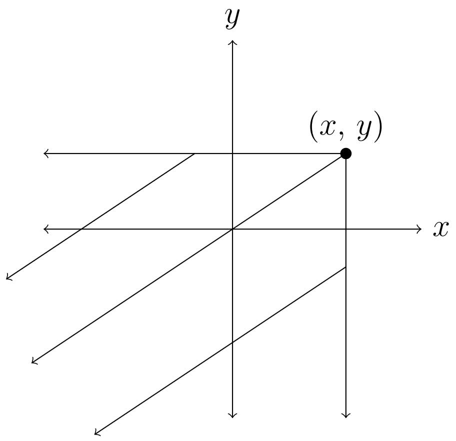


Figure 3.1: CDF is the probability that (X, Y) in the marked region

Theorem 3.1. *Let $\boldsymbol{X} = (X, Y)$ be a random vector with JCDF $F_{X,Y}(\cdot, \cdot)$. Then the CDF of X is given by $F_X(x) = \lim\limits_{y\to\infty} F_{X,Y}(x, y)$ for all $x \in \mathbb{R}$. Similarly, the CDF of Y is given by $F_Y(y) = \lim\limits_{x\to\infty} F_{X,Y}(x, y)$ for all $y \in \mathbb{R}$.*

Proof: Fix $x \in \mathbb{R}$. Let $\{y_n\}_{n\geq 1}$ be an increasing sequence of real numbers such that $y_n \to \infty$ as $n \to \infty$. Let us define

$$A_n = \{\omega \in \mathcal{S} : X(\omega) \leq x,\, Y(\omega) \leq y_n\}$$

for $n = 1, 2, 3, \ldots$. Clearly, $\{A_n\}_{n\geq 1}$ is an increasing sequence of events, and hence,

$$A = \lim_{n\to\infty} A_n = \cup_{n=1}^{\infty} A_n = \{\omega \in \mathcal{S} : X(\omega) \leq x\}.$$

Now,

$$\lim_{n\to\infty} F_{X,Y}(x, y_n) = \lim_{n\to\infty} P(A_n) = P\left(\lim_{n\to\infty} A_n\right) = P(A) = F_X(x).$$

This shows that for any increasing sequence of real numbers $\{y_n\}_{n\geq 1}$ with $y_n \to \infty$ as $n \to \infty$, $\lim_{n\to\infty} F_{X,Y}(x, y_n) = F_X(x)$. Thus, $\lim_{y\to\infty} F_{X,Y}(x, y) = F_X(x)$ for each fixed $x \in \mathbb{R}$.

Similarly, one can prove that $\lim_{x\to\infty} F_{X,Y}(x, y) = F_Y(y)$ for each fixed $y \in \mathbb{R}$. □

Remark 3.1. The previous theorem can be extended for more than two RVs. For example, let $\boldsymbol{X} = (X, Y, Z)$. In this case we can find CDFs of X, Y, and Z using the following formulas.

$$\begin{aligned}
F_X(x) &= \lim_{y\to\infty}\lim_{z\to\infty} F_{X,Y,Z}(x, y, z) = \lim_{z\to\infty}\lim_{y\to\infty} F_{X,Y,Z}(x, y, z) \quad \text{for all } x \in \mathbb{R},\\
F_Y(y) &= \lim_{x\to\infty}\lim_{z\to\infty} F_{X,Y,Z}(x, y, z) = \lim_{z\to\infty}\lim_{x\to\infty} F_{X,Y,Z}(x, y, z) \quad \text{for all } y \in \mathbb{R},\\
F_Z(z) &= \lim_{x\to\infty}\lim_{y\to\infty} F_{X,Y,Z}(x, y, z) = \lim_{y\to\infty}\lim_{x\to\infty} F_{X,Y,Z}(x, y, z) \quad \text{for all } z \in \mathbb{R}.
\end{aligned}$$

We can find the JCDFs of (X, Y), (X, Z), and (Y, Z) as

$$\begin{aligned}
F_{X,Y}(x, y) &= \lim_{z\to\infty} F_{X,Y,Z}(x, y, z) \quad \text{for all } (x, y) \in \mathbb{R}^2,\\
F_{X,Z}(x, z) &= \lim_{y\to\infty} F_{X,Y,Z}(x, y, z) \quad \text{for all } (x, z) \in \mathbb{R}^2,\\
F_{Y,Z}(y, z) &= \lim_{x\to\infty} F_{X,Y,Z}(x, y, z) \quad \text{for all } (y, z) \in \mathbb{R}^2.
\end{aligned}$$

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$. Let $A = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$. If we want to find the JCDF of $(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$, we need to take limit (tends to infinity) with respect to all the components that are not present in A. †

In the context of random vector, the JCDF of a subset is called marginal CDF. Thus, if $\boldsymbol{X} = (X, Y, Z)$, the CDF of X is called marginal CDF of X. Similarly the JCDF of (X, Y) is called marginal CDF of (X, Y).

Theorem 3.2 (Properties of JCDF)**.** *Let $\boldsymbol{X} = (X, Y)$ be a random vector with JCDF $F_{X,Y}(\cdot, \cdot)$. Then*

1. $\lim\limits_{x\to\infty} \lim\limits_{y\to\infty} F_{X,Y}(x, y) = 1$.

2. $\lim\limits_{x\to-\infty} F_{X,Y}(x, y) = 0$ *for all* $y \in \mathbb{R}$.

3. $\lim\limits_{y\to-\infty} F_{X,Y}(x, y) = 0$ *for all* $x \in \mathbb{R}$.

4. $F_{X,Y}(\cdot, \cdot)$ *is right continuous in each argument keeping other fixed.*

5. *For* $-\infty < a_1 < b_1 < \infty$ *and* $-\infty < a_2 < b_2 < \infty$,
$$F_{X,Y}(b_1, b_2) - F_{X,Y}(b_1, a_2) - F_{X,Y}(a_1, b_2) + F_{X,Y}(a_1, a_2) \geq 0.$$

Proof: The proof of this theorem is similar to that of Theorem 2.1 with standard modification for 2-dimensional functions. Therefore, the proof is skipped here. □

Though we are skipping the proof the previous theorem, let us make a comparison between the properties (that are presented in the Theorem 2.1) of CDF of a RV and that of JCDF of a 2-dimensional random vector. The Property 1 in the Theorem 2.1 states that $F_X(\cdot)$ is non-decreasing. This can be alternatively written as $F_X(x_2) - F_X(x_1) = P(x_1 < X \leq x_2) \geq 0$ for all $x_1 < x_2$. Thus, the non-decreasing property is a consequence of the fact that the probability that a RV is in an interval must be non-negative. Now, a natural extension of an interval in one-dimension is a rectangle in two-dimension. Thus, the equivalent property should be based on the fact that the probability that a 2-dimensional random vector in a rectangle must be non-negative. Let $a_1 < b_1$ and $a_2 < b_2$ be four real numbers. Then the points (a_1, a_2), (a_1, b_2), (b_1, a_2), and (b_1, b_2) forms vertices of the rectangle $(a_1, b_1] \times (a_2, b_2]$. Now,

$$\begin{aligned} P\left((X, Y) \in (a_1, b_1] \times (a_2, b_2]\right) &= P\left(a_1 < X \leq b_1, a_2 < Y \leq b_2\right) \\ &= F_{X,Y}(b_1, b_2) - F_{X,Y}(b_1, a_2) - F_{X,Y}(a_1, b_2) + F_{X,Y}(a_1, a_2). \end{aligned}$$

Thus, the fact that the probability that a random vector in a rectangle is non-negative gives the Property 5 of Theorem 3.2.

Property 2 of Theorem 2.1 states that $\lim_{x\to\infty} F_X(x) = 1$. This is intuitively tells that if we cover the whole $\mathbb{R}$, then the probability is one. Similarly, for a 2-dimensional random vector if the whole $\mathbb{R}^2$ is covered, the probability is one and we have Property 1 of Theorem 3.2.

Property 3 of Theorem 2.1 states that $\lim_{x\to-\infty} F_X(x) = 0$. Loosely speaking, $\{X \leq x\}$ becomes $\emptyset$ for x tends to $-\infty$. For 2-dimensional random vector, if one of the components tends to $-\infty$, the set becomes empty. If $x \to -\infty$, then $\{X \leq x\}$ becomes $\emptyset$, and hence, $\{X \leq x, Y \leq y\}$ becomes $\emptyset$. Similarly, as $y \to -\infty$, $\{X \leq x, Y \leq y\}$ becomes $\emptyset$. Thus, we have Properties 2 and 3 of Theorem 3.2. Note that for the first property of Theorem 3.2, both the components need to tend to ∞. However, for the Properties 2 and 3, if any of the components tends to $-\infty$ keeping other fixed, then JCDF tends to zero. Property 4 is a straight forward extension of Property 4 of the Theorem 2.1.

Theorem 3.3. *Let $G : \mathbb{R}^2 \to \mathbb{R}$ be a function satisfying conditions 1–5 of the Theorem 3.2. Then G is a JCDF of some 2-dimensional random vector.*

Proof: Proof of this theorem is out of scope of this course. □

This theorem can be used to check if a function is a JCDF or not. Theorems 3.2 and 3.3 can be extended for random vector having more than two components. However, writing the property (5) involves complicated expressions.

3.2 Discrete Random Vector

Definition 3.3 (Discrete Random Vector). *A random vector (X, Y) is said to have a discrete distribution if there exists an atmost countable set $S_{X,Y} \subset \mathbb{R}^2$ such that $P\left((X, Y) = (x, y)\right) = P\left(X = x, Y = y\right) > 0$ for all $(x, y) \in S_{X,Y}$ and $P\left((X, Y) \in S_{X,Y}\right) = 1$. $S_{X,Y}$ is called the support of (X, Y).*

Definition 3.4 (Joint PMF). *Let (X, Y) be a discrete random vector with support $S_{X,Y}$. Define a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ by*

$$f_{X,Y}(x, y) = \begin{cases} P(X = x, Y = y) & \text{if } (x, y) \in S_{X,Y} \\ 0 & \text{otherwise.} \end{cases}$$

The function $f_{X,Y}$ is called joint probability mass function (JPMF) of the discrete random vector (X, Y).

Note that discrete random vector is a straight forward extension of DRV. In case of DRV, we need to find an atmost countable set S_X in $\mathbb{R}$ such that $P\left(X \in S_X\right) = 1$. For a 2-dimensional discrete random vector $S_{X,Y}$ is atmost countable and a subset of $\mathbb{R}^2$ such that $P\left((X, Y) \in S_{X,Y}\right) = 1$. Similarly, the definition of JPMF is also a natural extension of PMF of a DRV. These definitions can be easily extended for more than two dimensional random vectors.

Theorem 3.4 (Properties of JPMF). *Let (X, Y) be a discrete random vector with JPMF $f_{X,Y}(\cdot, \cdot)$ and support $S_{X,Y}$. Then*

1. $f_{X,Y}(x, y) \geq 0$ *for* $(x, y) \in \mathbb{R}^2$.
2. $\displaystyle\sum_{(x,\, y) \in S_{X,\, Y}} f_{X,Y}(x, y) = 1$.

Proof: The proof of the theorem is straight forward form the definitions of discrete random vector and JPMF. □

Theorem 3.5. *If a function $g : \mathbb{R}^2 \to \mathbb{R}$ satisfy Properties 1 and 2 above for the atmost countable set $D = \{(x, y) \in \mathbb{R}^2 : g(x, y) > 0\}$ in place of $S_{X,Y}$, then g is JPMF of some 2-dimensional discrete random vector.*

Proof: The proof of this theorem is out of scope of this course. □

Theorem 3.5 can be used to check if a function is JPMF or not. Again, Theorems 3.4 and 3.5 can be extended for more than 2-dimensional discrete random vector.

Theorem 3.6 (Marginal PMF from JPMF). *Let (X, Y) be a discrete random vector with JPMF $f_{X,Y}(\cdot, \cdot)$ and support $S_{X,Y}$. Then X and Y are DRVs. The PMF of X is*

$$f_X(x) = \sum_{(x,\, y) \in S_{X,\,Y}} f_{X,Y}(x, y) \text{ for all fixed } x \in \mathbb{R}. \tag{3.1}$$

The PMF of Y is given by

$$f_Y(y) = \sum_{(x,\, y) \in S_{X,\,Y}} f_{X,Y}(x, y) \text{ for all fixed } y \in \mathbb{R}. \tag{3.2}$$

In this context, $f_X(\cdot)$ and $f_Y(\cdot)$ are called marginal PMF of X and marginal PMF of Y, respectively.

Proof: Define the set

$$D = \{x \in \mathbb{R} : (x, y) \in S_{X,Y} \text{ for some } y \in \mathbb{R}\}.$$

As, $S_{X,Y}$ is atmost countable, D is also atmost countable. Fix $x_0 \in D$. Consider $(x_0, y) \in S_{X,Y}$ and $(x_0, y') \in S_{X,Y}$ such that $y \neq y'$. Then the events

$$\{X = x_0, Y = y\} \text{ and } \{X = x_0, Y = y'\}$$

are disjoint. Now, using theorem of total probability (Theorem 1.17), $P(X = x_0)$ can be found by taking sum over all the points in $S_{X,Y}$ whose first component is x_0. Thus,

$$P(X = x_0) = \sum_{(x_0,\, y) \in S_{X,\,Y}} P(X = x, Y = y) = \sum_{(x_0,\, y) \in S_{X,\,Y}} f_{X,Y}(x, y),$$

and

$$\sum_{x \in D} P(X = x) = \sum_{x \in D} \sum_{(x,\, y) \in S_{X,\,Y}} f_{X,Y}(x, y) = \sum_{(x,\, y) \in S_{X,\,Y}} f_{X,Y}(x, y) = 1.$$

Hence, X is a DRV with PMF given by (3.1). Similarly we can prove that Y is also a DRV with PMF given by (3.2) □

Example 3.1. Let (X, Y) be a discrete random vector with JPMF

$$f(x, y) = \begin{cases} cy & \text{if } x = 1, 2, \ldots, n;\ y = 1, 2, \ldots, n \\ 0 & \text{otherwise,} \end{cases}$$

where c is a constant. We can find the value of c based on the properties of JPMF. If $f(\cdot,\, \cdot)$ have to be a JPMF, then $f(x,\, y) \geq 0$ for all $(x,\, y) \in \mathbb{R}^2$, which implies that $c \geq 0$. Also,

$$\sum_{x=1}^{n}\sum_{y=1}^{n} f(x,\, y) = 1 \implies c = \frac{2}{n^2(n+1)}.$$

Thus, the JPMF of $(X,\, Y)$ is given by

$$f(x,\, y) = \begin{cases} \frac{2y}{n^2(n+1)} & \text{if } x = 1,\, 2,\, \ldots,\, n;\; y = 1,\, 2,\, \ldots,\, n \\ 0 & \text{otherwise.} \end{cases}$$

We can also find the marginal PMF of X as follows: Fix $x \in \{1,\, 2,\, \ldots,\, n\}$. Then

$$P\left(X = x\right) = \sum_{(x,\, y) \in S_{X,\, Y}} f(x,\, y) = \sum_{y=1}^{n} cy = \frac{1}{n}.$$

Thus, the marginal PMF of X is given by

$$f_X(x) = \begin{cases} \frac{1}{n} & \text{if } x = 1,\, 2,\, \ldots,\, n \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, we can find the marginal PMF of Y. I leave it as an exercise. ||

Example 3.2. Let $(X,\, Y)$ be a discrete random vector with JPMF

$$f(x,\, y) = \begin{cases} cy & \text{if } x = 1,\, 2,\, \ldots,\, n;\; y = 1,\, 2,\, \ldots,\, n;\; x \leq y \\ 0 & \text{otherwise,} \end{cases}$$

where c is a constant. Note that the function $f(\cdot,\, \cdot)$ is almost similar to that of the previous example. Only difference is in the sets where the functions are strictly positive. In the previous example, f was positive on $\{1,\, 2,\, \ldots,\, n\} \times \{1,\, 2,\, \ldots,\, n\}$. In the current example the set is $\{(x,\, y) \in \mathbb{R}^2 : x = 1,\, 2,\, \ldots,\, n;\; y = 1,\, 2,\, \ldots,\, n;\; x \leq y\}$. However, this changes the probability distribution completely. We will see that the marginal PMFs are also different. Hence, support is an important issue.

The constant c is positive and can be found as follows:

$$\sum_{(x,\, y) \in S_{X,\, Y}} f(x,\, y) = 1 \implies \sum_{y=1}^{n}\sum_{x=1}^{y} cy = 1 \implies c = \frac{6}{n(n+1)(2n+1)}.$$

Please note the range of the summations. Thus, the JPMF of $(X,\, Y)$ is given by

$$f(x,\, y) = \begin{cases} \frac{6y}{n(n+1)(2n+1)} & \text{if } x = 1,\, 2,\, \ldots,\, n;\; y = 1,\, 2,\, \ldots,\, n;\; x \leq y \\ 0 & \text{otherwise} \end{cases}$$

The marginal PMF of X can be found as follows: For $x \in \{1,\, 2,\, \ldots,\, n\}$,

$$P\left(X = x\right) = \sum_{y=x}^{n} cy = \frac{3(n+x)(n-x+1)}{n(n+1)(2n+1)}.$$

Please note the range of the summation above. Thus, the marginal PMF of X is given by

$$f_X(x) = \begin{cases} \frac{3(n+x)(n-x+1)}{n(n+1)(2n+1)} & \text{if } x = 1, 2, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

The marginal PMF of Y can also be found similarly and is given by

$$f_Y(y) = \begin{cases} \frac{6y^2}{n(n+1)(2n+1)} & \text{if } y = 1, 2, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

||

3.3 Continuous Random Vector

Definition 3.5 (Continuous Random Vector). *A random vector (X, Y) is said to have a continuous distribution if there exists a non-negative integrable function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ such that*

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t, s)dsdt$$

for all $(x, y) \in \mathbb{R}^2$. The function $f_{X,Y}$ is called the joint probability density function (JPDF) of (X, Y). The set $S_{X,Y} = \{(x, y) \in \mathbb{R}^2 : f_{X,Y}(x, y) > 0\}$ is called the support of (X, Y).

Again, the continuous random vector is a natural extension of CRV. The JPMF exists only for discrete random vector and JPDF exits for continuous random vector.

Theorem 3.7 (Properties of JPDF). *Let (X, Y) be a continuous random vector with JPDF $f_{X,Y}(\cdot, \cdot)$. Then*

1. *$f_{X,Y}(x, y) \geq 0$ for $(x, y) \in \mathbb{R}^2$.*
2. $\displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1.$

Proof: The proof of the Property 1 is straight forward from the definition of continuous random vector. For the Property 2,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = \lim_{A\to\infty} \lim_{B\to\infty} \int_{-\infty}^{A} \int_{-\infty}^{B} f_{X,Y}(x, y)dxdy = \lim_{A\to\infty} \lim_{B\to\infty} F_{X,Y}(A, B) = 1.$$

□

Theorem 3.8. *If a function $g : \mathbb{R}^2 \to \mathbb{R}$ satisfy Properties 1 and 2 of the Theorems 3.7, then $g(\cdot, \cdot)$ is JPDF of some 2-dimensional continuous random vector.*

Proof: The proof of this theorem is out of scope of this course. □

Theorem 3.9 (Marginal PDF from JPDF). *Let (X, Y) be a continuous random vector with JPDF $f_{X,Y}(\cdot, \cdot)$. Then X and Y are CRVs. The PDF of X is given by*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \quad \textit{for all fixed } x \in \mathbb{R}. \tag{3.3}$$

The PDF of Y *is given by*

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,\, y)dx \quad \text{for all fixed } y \in \mathbb{R}. \tag{3.4}$$

In the context of continuous random vector, $f_X(\cdot)$ *and* $f_Y(\cdot)$ *are called marginal PDF of* X *and marginal PDF of* Y*, respectively.*

Proof: For $x \in \mathbb{R}$,

$$\begin{aligned}
F_X(x) &= \lim_{y\to\infty} F_{X,Y}(x,\, y) \\
&= \lim_{y\to\infty} \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)dtds \\
&= \int_{-\infty}^{x} \left\{ \lim_{y\to\infty} \int_{-\infty}^{y} f_{X,Y}(s,\, t)dt \right\} ds \\
&= \int_{-\infty}^{x} \left\{ \int_{-\infty}^{\infty} f_{X,Y}(s,\, t)dt \right\} ds \\
&= \int_{-\infty}^{x} g(s)ds,
\end{aligned}$$

where $g(s) = \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(s,\, t)dt$. The third equality holds true as $f_{X,Y}(x,\, y) \geq 0$ for all $(x,\, y) \in \mathbb{R}^2$. Thus, X is a CRV with PDF as given in (3.3). Similarly, we can prove that Y is also CRV with PDF given in (3.4). □

Example 3.3. Let $(X,\, Y)$ be a CRV with JPDF

$$f(x,\, y) = \begin{cases} ce^{-(2x+3y)} & \text{if } 0 < x < y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where c is a constant. Clearly, $c > 0$ as $f_{X,Y}(x,\, y) \geq 0$ for all $(x,\, y) \in \mathbb{R}^2$. The value of c can be found as follows:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,\, y)dydx = 1 \implies c \int_{0}^{\infty} \int_{x}^{\infty} e^{-(2x+3y)}dydx = 1 \implies c = 15.$$

Note the range of integration. We can find the marginal PDF of X as follows: For $x \leq 0$,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,\, y)dy = 0,$$

as the integrand is zero for all $x \leq 0$. For $x > 0$,

$$f_X(x) = \int_{-\infty}^{x} f(x,\, y)dy + \int_{x}^{\infty} f(x,\, y)dy = \int_{-\infty}^{\infty} f(x,\, y)dy = 15 \int_{x}^{\infty} e^{-(2x+3y)}dy = 5e^{-5x},$$

as $f(x,\, y) = 0$ for $y < x$. Thus, the marginal PDF of X is given by

$$f_X(x) = \begin{cases} 5e^{-5x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $X \sim Exp(5)$. Similarly, the marginal PDF of Y can be calculated and is given by

$$f_Y(y) = \begin{cases} \frac{15}{2}e^{-3y}\left(1 - e^{-2y}\right) & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

||

3.4 Expectation of Function of Random Vector

Definition 3.6 (Expectation of Function of Discrete Random Vector)**.** *Let $(X,\, Y)$ be a discrete random vector with JPMF $f_{X,Y}\left(\cdot,\, \cdot\right)$ and support $S_{X,Y}$. Let $h:\mathbb{R}^2 \to \mathbb{R}$. Then the expectation of $h(X,\, Y)$ is defined by*

$$E\left(h(X,\, Y)\right) = \sum_{(x,\, y)\in S_{X,\, Y}} h(x,\, y) f_{X,Y}(x,\, y),$$

provided $\displaystyle\sum_{(x,\, y)\in S_{X,\, Y}} |h(x,\, y)| f_{X,Y}(x,\, y) < \infty.$

Definition 3.7 (Expectation of Function of Continuous Random Vector)**.** *Let $(X,\, Y)$ be a continuous random vector with JPDF $f_{X,Y}(\cdot,\, \cdot)$. Let $h:\mathbb{R}^2 \to \mathbb{R}$. Then the expectation of $h(X,\, Y)$ is defined by*

$$E\left(h(X,\, Y)\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,\, y) f_{X,Y}(x,\, y) dxdy,$$

provided $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |h(x,\, y)| f_{X,Y}(x,\, y) dxdy < \infty.$

Theorem 3.10 (Linearity Property of Expectation)**.** *Let $\boldsymbol{X} = (X_1,\, X_2,\, \ldots,\, X_n)$ be either a discrete random vector or a continuous random vector. Then*

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$$

where $a_i \in \mathbb{R}$ is a constant for all $i = 1,\, 2,\, \ldots,\, n$. Here we assume that all the expectations exist.

Proof: We will prove the theorem for continuous random vector $\boldsymbol{X} = (X,\, Y)$. The proof for general value of n is similar. Also, the proof for discrete random vector can be written easily by replacing integration sign by summation sign. Let f be the JPDF of $\boldsymbol{X}$. Then

$$\begin{aligned}
E\left(a_1 X + a_2 Y\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(a_1 x + a_2 y\right) f_{X,Y}(x,\, y) dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a_1 x f_{X,Y}(x,\, y) dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a_2 y f_{X,Y}(x,\, y) dxdy \\
&= a_1 \int_{-\infty}^{\infty} x \left\{\int_{-\infty}^{\infty} f_{X,Y}(x,\, y) dy\right\} dx + a_2 \int_{-\infty}^{\infty} y \left\{\int_{-\infty}^{\infty} f_{X,Y}(x,\, y) dx\right\} dy \\
&= a_1 \int_{-\infty}^{\infty} x f_X(x) dx + a_2 \int_{-\infty}^{\infty} y f_Y(y) dy \\
&= a_1 E\left(X\right) + a_2 E\left(Y\right).
\end{aligned}$$

□

The previous theorem tells us that we can compute expectation of a linear combination of a random vectors by computing the expectations of individual components of the random vector and then taking the linear combination. Note that the left hand side involves the joint distribution and an n-dimensional integration (for continuous random vector) or an n-dimensional summation (for discrete random vector). However, the right hand side involves n one-dimensional integrations or n one-dimensional summations. Also, we need the marginal distributions (not joint distribution) to compute the right hand side. Sometimes it is much easier to compute n one-dimensional integrations compared to a single n-dimensional integration. Same is true for summations. Also, in many problems, it is easier to obtain the marginal distributions than to obtain joint distribution. The following example illustrate this.

Example 3.4. At a party n men throw their hats into the center of a room. The hats are mixed up and each man randomly selects one. Suppose that we want to calculate the expected number of men who selects their own hat. Let X denote the number of men who selects their own hat. We are interested to find $E(X)$. To compute $E(X)$ directly, we need the distribution of X. It is clear that X should be a DRV with support $S_X = \{0, 1, \ldots, n-2, n\}$. Note that $n - 1 \notin S_X$ *(why?)*. It is quite easy to find the values $P(X = k)$ for $k = 0$ and n. However, it is quite difficult to find $P\left(X = k\right)$ for other values of $k \in S_X$. Thus, it becomes a difficult problem if we try to compute $E(X)$ directly.

Let us try to solve it by converting the problem into a multidimensional problem. For $i = 1, 2, \ldots, n$, let us define the RV

$$X_i = \begin{cases} 1 & \text{if } i\text{th person takes his own hat} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, $X = X_1 + X_2 + \ldots + X_n$. Thus, using the previous theorem, $E(X) = E(X_1) + \ldots + E(X_n)$. Now, we need to compute $E(X_i)$ for all $i = 1, 2, \ldots, n$. Note that $P\left(X_i = 1\right) = \frac{1}{n}$ for all $i = 1, 2, \ldots, n$. Hence, $E(X_i) = \frac{1}{n}$ for all $i = 1, 2, \ldots, n$ implies $E(X) = 1$. Thus, on an average only one person takes his own hat, and it does not depend on n, the number of persons present in the game. ||

3.5 Some Useful Remarks

Remark 3.2. In Theorem 3.6, we have seen that if (X, Y) is a discrete random vector then X and Y are DRVs. On the other hand, suppose that X and Y are DRVs with supports S_X and S_Y, respectively. Suppose that $S = S_X \times S_Y$. Clearly, S is atmost countable. Then

$$\begin{aligned} P\left((X, Y) \in S\right) &= \sum_{(x,\, y) \in S} P\left(X = x,\, Y = y\right) \\ &= \sum_{x \in S_X} \left\{ \sum_{y \in S_Y} P\left(X = x,\, Y = y\right) \right\} \\ &= \sum_{x \in S_X} P\left(X = x\right), \quad \text{using theorem of total probability} \\ &= 1. \end{aligned}$$

Let $T = \{(x, y) \in \mathbb{R}^2 : P\left(X = x, Y = y\right) > 0\}$. Then $T \subseteq S$, and hence, T is atmost countable. Also, $P\left((X, Y) \in T\right) = 1$. Thus, (X, Y) is discrete random vector. This discussion shows that (X, Y) is discrete random vector if and only if X and Y are DRVs. †

Remark 3.3. If $(X,\, Y)$ is continuous random vector, then

$$P\left((X,\, Y) \in A)\right) = \int\int_{(x,\, y) \in A} f_{X,Y}(x,\, y)dxdy,$$

for all $A \subseteq \mathbb{R}^2$ such that the integration is possible. This statement can be seen as an extension of the fact that if X is a CRV with PDF $f(\cdot)$, then $P\left(X \in B\right) = \int_B f(x)dx$. However, the mathematical proof is out of the scope of this course. †

Remark 3.4. In Theorem 3.9, we have seen that if $(X,\, Y)$ is continuous random vector, then X and Y are continuous random variables. However, the converse, in general, is not true. Thus, $(X,\, Y)$ may not be a continuous random vector even if X and Y are CRVs. Consider the following example in this regards. Let X be a CRV. Suppose that $Y = X$. Then X and Y are CRVs. It is clear that $P\left(X = Y\right) = 1$. Now, if possible, assume that $(X,\, Y)$ is a continuous random vector. Thus, $(X,\, Y)$ has a JPDF, say $f(\cdot,\, \cdot)$. Then

$$P\left(X = Y\right) = \int\int_{x=y} f(x, y)dxdy = 0.$$

The last equality is true as a double integral $\iint_B g(x,\, y)dxdy$ can be interpreted as the volume under the function $g(\cdot,\, \cdot)$ over the set B. As the area of the set $\{(x,\, y) \in \mathbb{R}^2 : x = y\}$ is zero, the volume is also zero. This is a contradiction to the fact that $P\left(X = Y\right) = 1$. Thus, our assumption is wrong and $(X,\, Y)$ is not a continuous random vector.

In general, if there exists a set $A \subset \mathbb{R}^2$ whose area is zero and $P\left((X,\, Y) \in A\right) > 0$, then $(X,\, Y)$ does not have a JPDF, and hence, $(X,\, Y)$ is not a continuous random vector. †

Remark 3.5. In Theorems 3.6 and 3.9, we have seen that the marginal distributions can be recovered form the joint distribution. However, the converse, in general, is not true. Let us illustrate it using the following example. †

Example 3.5. Let $f(\cdot)$ and $g(\cdot)$ be two PDFs and $F(\cdot)$ and $G(\cdot)$ be the corresponding CDFs, respectively. Define, for $-1 < \alpha < 1$,

$$h(x,\, y) = f(x)g(y)\left\{1 + \alpha(1 - 2F(x))(1 - 2G(y))\right\}.$$

First, we will show that $h(\cdot,\, \cdot)$ is a JPDF of a two-dimensional random vector. As $0 \leq F(x) \leq 1$, $-1 \leq 1 - 2F(x) \leq 1$. Similarly $-1 \leq 1 - 2G(y) \leq 1$. Hence, for all $(x,\, y) \in \mathbb{R}^2$,

$$-|\alpha| \leq \alpha\left(1 - 2F(x)\right)\left(1 - 2G(y)\right) \leq |\alpha| \implies 1 + \alpha\left(1 - 2F(x)\right)\left(1 - 2G(y)\right) \geq 0.$$

As $f(x) \geq 0$ and $g(y) \geq 0$, $h(x,\, y) \geq 0$ for all $(x,\, y) \in \mathbb{R}^2$. Also,

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,\, y)dxdy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)g(y)\left\{1 + \alpha(1 - 2F(x))(1 - 2G(y))\right\}dxdy \\
&= 1 + \alpha\left(\int_{\infty}^{\infty} f(x)\left(1 - 2F(x)\right)dx\right)\left(\int_{-\infty}^{\infty} g(y)\left(1 - 2G(y)\right)dy\right) \\
&= 1 + \alpha\left(\int_{\infty}^{\infty} \left(1 - 2F(x)\right)dF(x)\right)\left(\int_{-\infty}^{\infty} \left(1 - 2G(y)\right)dG(y)\right) \\
&= 1.
\end{aligned}$$

Thus, $h(\cdot, \cdot)$ is a JPDF of a 2-dimensional continuous random vector, say (X, Y). Let us try to find the marginal PDFs of X and Y. The marginal PDF of X is

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} h(x, y)dy \\ &= f(x)\int_{-\infty}^{\infty} g(y)dy + \alpha f(x)\left(1 - 2F(x)\right)\int_{-\infty}^{\infty} g(y)\left(1 - 2G(y)\right)dy \\ &= f(x). \end{aligned}$$

Similarly the marginal PDF of Y is $g(\cdot)$. Thus, the marginal PDFs of X and Y does not depend on α. However, the JPDF depends on the value of α. For different values of α, we have different JPDF, but the marginals remain same. Hence, given the marginal distributions, in general, we cannot construct the joint distribution. $||$

3.6 Independent Random Variables

Definition 3.8 (Independent RVs). *The random variables $X_1, X_2, \ldots, X_n$ are said to be independent if*

$$F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i),$$

for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, where $F_{X_i}(\cdot)$ is the marginal CDF of X_i.

Thus, two RVs X and Y are independent if and only if the events $E_x = \{X \leq x\}$ and $F_y = \{Y \leq y\}$ are independent for all $(x, y) \in \mathbb{R}^2$. For discrete random vector, an equivalent definition of independence can be given in terms of JPMF and marginal PMFs. Similarly, for continuous random vector, an equivalent definition of independence can be given in terms of JPDF and marginal PDFs. Next, we will give these two alternative definitions, however we will not prove the equivalence of the respective definitions. Nonetheless, these alternative definitions are very handy in many applications.

Definition 3.9 (Alternative Definition for Discrete Random Vector). *Let $(X_1, X_2, \ldots, X_n)$ be a discrete random vector with JPMF $f_{X_1, X_2, \ldots, X_n}(\cdot, \ldots, \cdot)$. Then $X_1, X_2, \ldots, X_n$ are said to be independent if*

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, where $f_{X_i}(\cdot)$ is the marginal PMF of X_i, $i = 1, 2, \ldots, n$.

Definition 3.10 (Alternative Definition for Continuous Random Vector). *Let $(X_1, X_2, \ldots, X_n)$ be a continuous random vector with JPDF $f_{X_1, X_2, \ldots, X_n}(\cdot, \ldots, \cdot)$. Then $X_1, X_2, \ldots, X_n$ are said to be independent if*

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, where $f_{X_i}(\cdot)$ is the marginal PDF of X_i, $i = 1, 2, \ldots, n$.

In the last section, we have pointed out that, in general, the joint distribution cannot be recovered from the marginal distributions. However, if $X_1, X_2, \ldots, X_n$ are independent RVs and if we know the marginal distributions of X_i for all $i = 1, 2, \ldots, n$, then we can write $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$ for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Thus, we can recover the joint distribution from the marginal distributions if the RVs are known to be independent. Moreover, if $X_1, X_2, \ldots, X_n$ are independent CRVs, then $(X_1, \ldots, X_n)$ is a continuous random vector.

Theorem 3.11. *If X and Y are independent, then*

$$E\left(g(X)h(Y)\right) = E\left(g(X)\right) E\left(h(Y)\right),$$

provided all the expectations exist.

Proof: We will prove it for continuous random vector (X, Y). For discrete random vector, it can be proved by replacing the integration sign by summation sign.

$$\begin{aligned} E\left(g(X)h(Y)\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_{X,Y}(x, y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy, \quad \text{as } X \text{ and } Y \text{ are independent} \\ &= \left(\int_{-\infty}^{\infty} g(x)f_X(x)dx\right)\left(\int_{-\infty}^{\infty} h(y)f_Y(y)dy\right) \\ &= E\left(g(X)\right)E\left(h(Y)\right). \end{aligned}$$

□

3.7 Covariance and Correlation Coefficient

Definition 3.11 (Covariance)**.** *The covariance of two random variables X and Y is defined by*

$$Cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right] = E(XY) - E(X)E(Y).$$

Definition 3.12 (Correlation Coefficient)**.** *The correlation coefficient of X and Y is defined by*

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

Remark 3.6. Let X and Y be independent, then

$$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0.$$

We get the second equality using the previous theorem. Thus, $\rho(X, Y) = 0$. However, the converse is not true in general. That means that there exists dependent RVs X and Y such that $Cov(X, Y) = 0$. Let us consider the following example in this regards. †

Example 3.6. Let $X \sim N(0, 1)$ and $Y = X^2$. Then

$$Cov(X, Y) = E\left(XY\right) - E(X)E(Y) = E\left(X^3\right) - E(X)E(Y).$$

It is easy to see that $E(X) = 0$ and $E(X^3) = 0$. Hence, $Cov(X, Y) = 0$. Now, $P\left(X \leq -5\right) = \Phi(-5) \neq 0$ and $P\left(Y \leq 1\right) = P\left(-1 \leq X \leq 1\right) = 2\Phi(1) - 1 \neq 0$. However, $P\left(X \leq -5, Y \leq 1\right) = 0$. Thus, X and Y are not independent. ||

Theorem 3.12. *Let X, Y, Z, $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ be RVs such that all the necessary expectations for the followings exist. Then*

1. $Cov(X, X) = Var(X)$.
2. $Cov(X, Y) = Cov(Y, X)$.
3. $Cov(aX, Y) = a\,Cov(X, Y)$ *for a real constant* a.
4. $Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$.
5. $Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j)$ *for real constants* $a_1, a_2, \ldots, a_n$ *and* $b_1, b_2, \ldots, b_m$.
6. $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, Y_j)$.
7. *If* X_i*'s are independent, then* $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$.

Proof:

1. Straight forward form the definition.
2. Straight forward form the definition.
3.

$$\begin{aligned} Cov\,(aX, Y) &= E\,((aX - E\,(aX))\,(Y - E(Y))) \\ &= aE\,((X - E(X))\,(Y - E(Y))) \\ &= a\,Cov(X, Y). \end{aligned}$$

4. Straight forward from the definition.
5. Combining 2, 3, and 4, we can prove it.
6. Combining 1 and 5, this proof is trivial.
7. Using Remark 3.6, it can be readily obtained from 5.

□

Theorem 3.13. $|\rho(X, Y)| \leq 1$ *provided it exists.*

Proof: Note that for any $\lambda \in \mathbb{R}$,

$$Var(X + \lambda Y) \geq 0 \implies \lambda^2 Var(Y) + 2\lambda Cov(X, Y) + Var(X) \geq 0.$$

That means that the quadratic equation $\lambda^2 Var(Y) + 2\lambda Cov(X, Y) + Var(X) = 0$ either has one real solution or no real solutions. Hence,

$$4\,(Cov\,(X, Y))^2 - 4Var(X)Var(Y) \leq 0 \implies |\rho(X, Y)| \leq 1.$$

□

3.8 Transformation Techniques

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a random vector and $g : \mathbb{R}^n \to \mathbb{R}^m$. Clearly, $\boldsymbol{Y} = g(\boldsymbol{X})$ is a m-dimensional random vector. In this section, we will discuss different methods to find the distribution of the random vector $\boldsymbol{Y} = g(\boldsymbol{X})$. Like the previous chapter, there are mainly three techniques to obtain the distribution of $\boldsymbol{Y} = g(\boldsymbol{X})$.

3.8.1 Technique 1

In Technique 1, we try to find the JCDF of $\boldsymbol{Y} = g(\boldsymbol{X})$ given the distribution of $\boldsymbol{X}$. As before, we will discuss this technique using examples.

Example 3.7. Let X_1 and X_2 be identically and independently distributed (*i.i.d.*) $U(0, 1)$ random variables. Suppose we want to find the CDF of $Y = X_1 + X_2$. Now,

$$F_Y(y) = P\left(Y \leq y\right) = P\left(X_1 + X_2 \leq y\right) = \int\int_{x_1+x_2\leq y} f_{X_1,X_2}(x_1, x_2)dx_1dx_2. \tag{3.5}$$

As $X_1 \sim U(0, 1)$, $X_2 \sim U(0, 1)$ and X_1 and X_2 are independent RVs, the JPDF of (X_1, X_2) is given by

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} 1 & \text{if } 0 < x_1 < 1,\, 0 < x_2 < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the JPDF of (X_1, X_2) is positive only on the unit square $(0, 1) \times (0, 1)$, which is indicated by gray shade in Figure 3.2. Now, to compute the integration in (3.5), we need to consider the following cases.

For $y < 0$, consider the Figure 3.2a. As the integrand in (3.5) is zero over the region $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$ for $y < 0$,

$$F_Y(y) = 0.$$

For $0 \leq y < 1$, consider the Figure 3.2b. The integrand is positive only on the shaded region in the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$. Therefore,

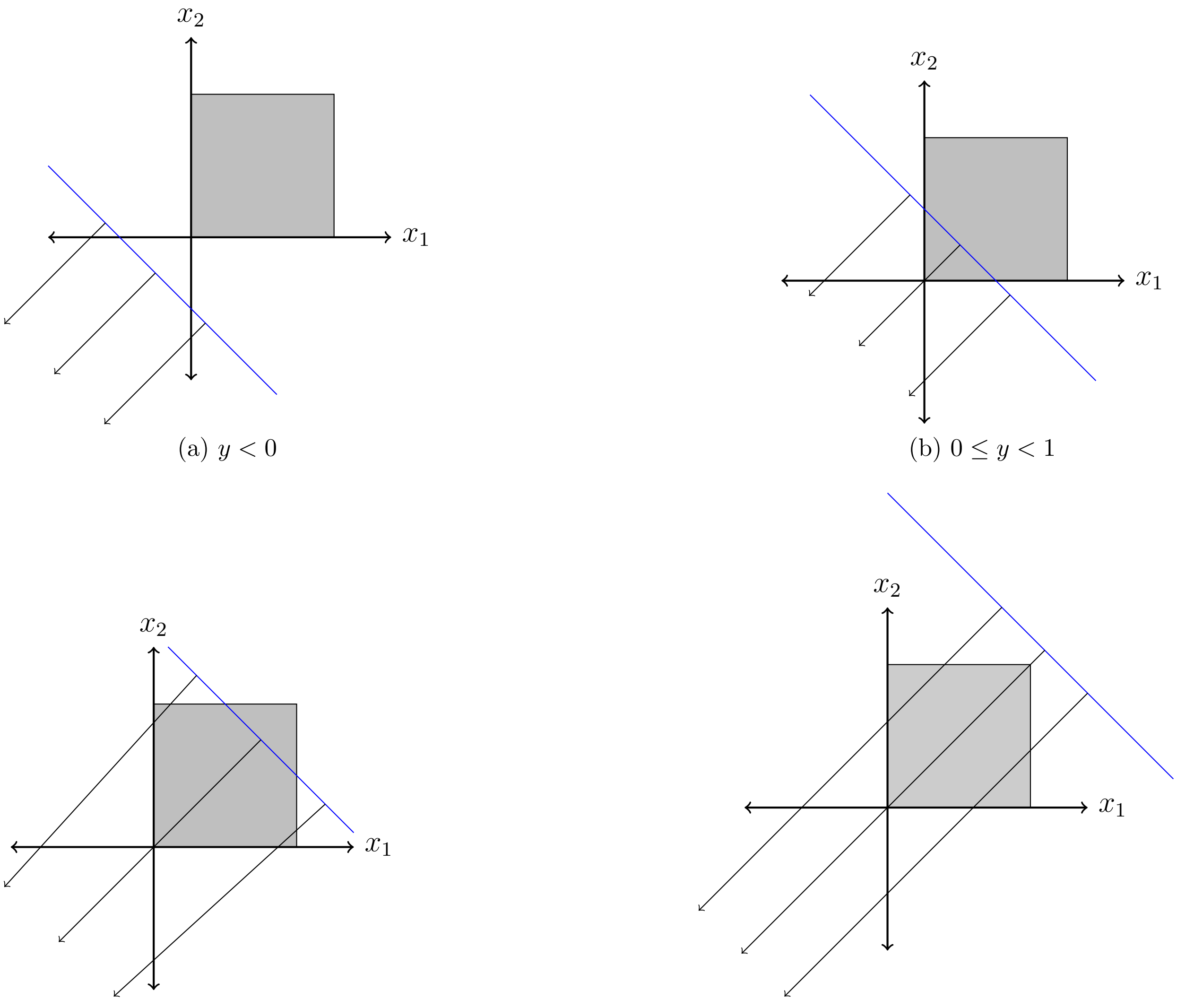
$$F_Y(y) = \int_0^y \int_0^{y-x_2} dx_1dx_2 = \frac{1}{2}y^2.$$

For $1 \leq y < 2$, consider the Figure 3.2c. The integrand is positive only on the shaded region in the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$. Therefore,

$$F_Y(y) = 1 - \int_{y-1}^1 \int_{y-x_2}^1 dx_1dx_2 = 1 - \frac{1}{2}(2-y)^2.$$

For $y \geq 2$, consider the Figure 3.2d. The integrand is positive on the shaded region in the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$ and the square $(0, 1) \times (0, 1)$ is completely inside the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$. Therefore,

$$F_Y(y) = 1.$$



(a) $y < 0$

(b) $0 \leq y < 1$

(c) $1 \leq y < 2$

(d) $y \geq 2$

Figure 3.2: Plot for Example 3.7

Thus, the CDF of $Y = X_1 + X_2$ is given by

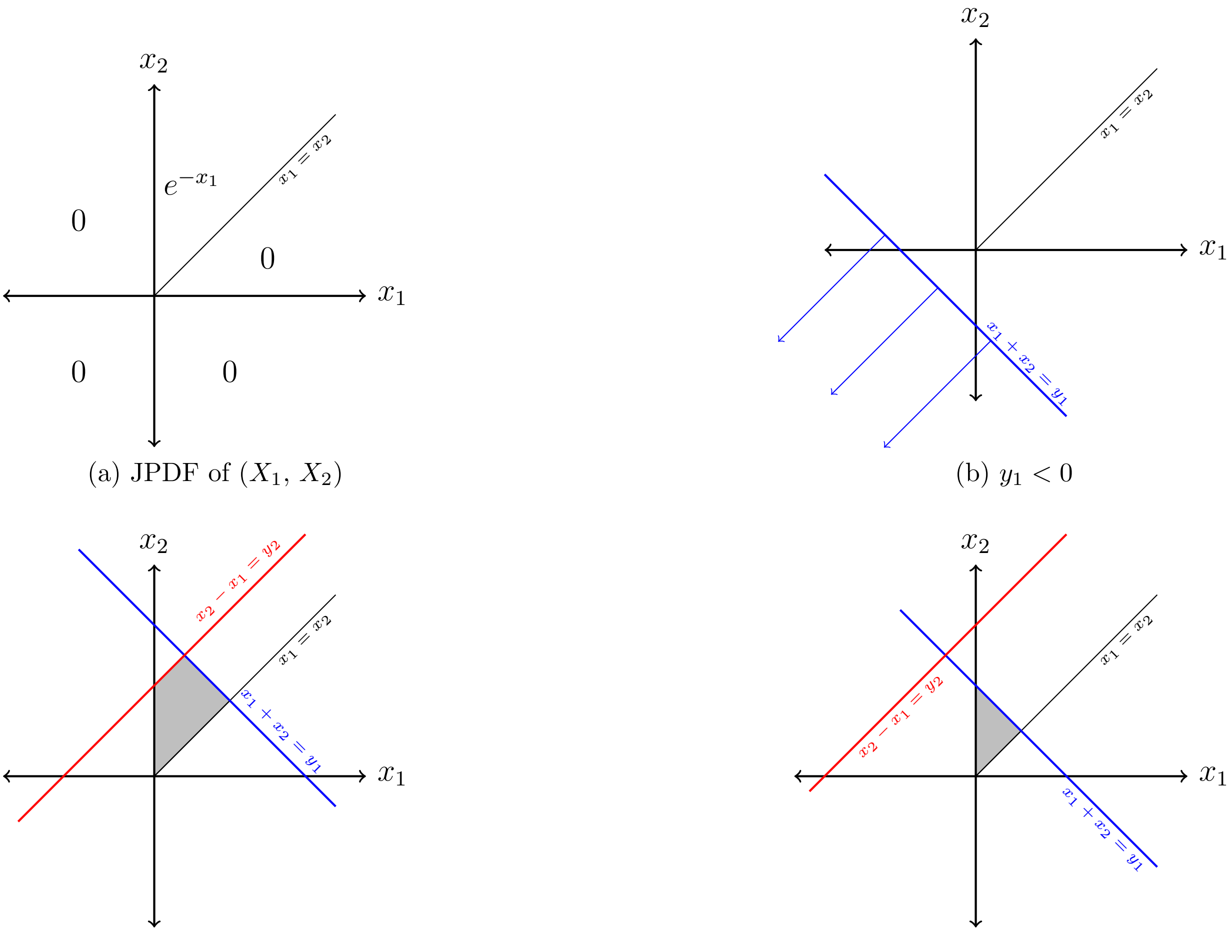
$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2}y^2 & \text{if } 0 \leq y < 1 \\ 1 - \frac{1}{2}(2-y)^2 & \text{if } 1 \leq y < 2 \\ 1 & \text{if } y \geq 2. \end{cases}$$

It can be shown that Y is a CRV *(why?)*. $||$

Example 3.8. Let the JPDF of (X_1, X_2) be given by

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} e^{-x_1} & \text{if } 0 < x_1 < x_2 < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that we want to find the JCDF of $Y_1 = X_1 + X_2$ and $Y_2 = X_2 - X_1$. Note that the JPDF of (X_1, X_2) is positive only on the set $S_{X_1, X_2} = \{(x_1, x_2) \in \mathbb{R}^2 : 0 < x_1 < x_2 < \infty\}$.



(a) JPDF of $(X_1,\, X_2)$

(b) $y_1 < 0$

(c) $0 < y_2 < y_1 < \infty$

(d) $0 < y_1 < y_2 < \infty$

Figure 3.3: Plot for Example 3.8

See Figure 3.3a. Now, let $A_{y_1, y_2} = \{(x_1,\, x_2) \in \mathbb{R} : x_1 + x_2 \le y_1,\, x_2 - x_1 \le y_2\}$. Then

$$F_{Y_1, Y_2}(y_1,\, y_2) = P\left(X_1 + X_2 \le y_1,\, X_2 - X_1 \le y_2\right) = \int\int_{A_{y_1,\, y_2}} f_{X_1,\, X_2}(x_1,\, x_2) dx_2 dx_1. \quad (3.6)$$

Suppose that $y_1 < 0$. Then $F_{Y_1}(y_1) = 0$. See the Figure 3.3b. As $F_{Y_1, Y_2}(y_1,\, y_2) \le \min\{F_{Y_1}(y_1),\, F_{Y_2}(y_2)\}$, $F_{Y_1, Y_2}(y_1,\, y_2) = 0$ for $y_1 < 0$. Similarly, $F_{Y_1, Y_2}(y_1,\, y_2) = 0$ for $y_2 < 0$. For $0 < y_2 < y_1 < \infty$, $A_{y_1, y_2} \cap S_{X_1,\, X_2}$ is the shaded region of the Figure 3.3c. Therefore,

$$\begin{aligned} F_{Y_1, Y_2}(y_1,\, y_2) &= \int_0^{\frac{y_1 - y_2}{2}} \int_{x_1}^{x_1 + y_1} e^{-x_1} dx_2 dx_1 + \int_{\frac{y_1 - y_2}{2}}^{\frac{y_1}{2}} \int_{x_1}^{y_1 - x_1} e^{-x_1} dx_2 dx_1 \\ &= y_1 + e^{-\frac{y_1}{2}} - (y_1 - y_2 + 2) e^{-\frac{y_1 - y_2}{2}}. \end{aligned}$$

For $0 < y_1 < y_2 < \infty$, $A_{y_1, y_2} \cap S_{X_1,\, X_2}$ is indicated by the shaded region in the Figure 3.3d. Therefore,

$$F_{Y_1, Y_2}(y_1,\, y_2) = \int_0^{\frac{y_1}{2}} \int_{x_1}^{y_1 - x_1} e^{-x_1} dx_2 dx_1 = y_1 + 2e^{-\frac{y_1}{2}} - 2.$$

Thus, the JCDF of $(Y_1,\, Y_2) = (X_1 + X_2,\, X_2 - X_1)$ is given by

$$F_{Y_1,\,Y_2}(y_1,\, y_2) = \begin{cases} 0 & \text{if } y_1 < 0 \text{ or } y_2 < 0 \\ y_1 + e^{-\frac{y_1}{2}} - (y_1 - y_2 + 2)e^{-\frac{y_1 - y_2}{2}} & \text{if } 0 < y_2 \le y_1 < \infty \\ y_1 + 2e^{-\frac{y_1}{2}} - 2 & \text{if } 0 < y_1 < y_2 < \infty. \end{cases}$$

It can be shown that $(Y_1,\, Y_2)$ is a continuous random vector. This is easy and therefore left as an exercise. ||

3.8.2 Technique 2

Like the previous chapter, this technique is based on two theorems, one for discrete random vector and another for continuous random vector.

Theorem 3.14. *Let $\boldsymbol{X} = (X_1,\, X_2,\, \ldots,\, X_n)$ be a discrete random vector with JPMF $f_{\boldsymbol{X}}$ and support $S_{\boldsymbol{X}}$. Let $g_i : \mathbb{R}^n \to \mathbb{R}$ for all $i = 1,\, 2,\, \ldots,\, k$. Let $Y_i = g_i(\boldsymbol{X})$ for $i = 1,\, 2,\, \ldots,\, k$. Then $\boldsymbol{Y} = (Y_1, \ldots,\, Y_k)$ is a discrete random vector with JPMF*

$$f_{\boldsymbol{Y}}(y_1,\, \ldots,\, y_k) = \begin{cases} \displaystyle\sum_{\boldsymbol{x} \in A_{\boldsymbol{y}}} f_{\boldsymbol{X}}(\boldsymbol{x}) & \textit{if } (y_1,\, \ldots,\, y_k) \in S_{\boldsymbol{Y}} \\ 0 & \textit{otherwise,} \end{cases}$$

where $A_{\boldsymbol{y}} = \{\boldsymbol{x} \in S_{\boldsymbol{X}} : g_i(\boldsymbol{x}) = y_i,\, i = 1,\, \ldots,\, k\}$ and $S_{\boldsymbol{Y}} = \{(g_1(\boldsymbol{x}),\, \ldots,\, g_k(\boldsymbol{x})) : \boldsymbol{x} \in S_{\boldsymbol{X}}\}$.

Proof: The proof of the theorem is similar to that of Theorem 2.7. □

Example 3.9. Let $X_1 \sim Poi(\lambda_1)$ and $X_2 \sim Poi(\lambda_2)$. Also, assume that X_1 and X_2 are independent. Then $Y = X_1 + X_2 \sim Poi(\lambda_1 + \lambda_2)$. To see it, we can apply Theorem 3.14. First note that the JPMF if $(X_1,\, X_2)$ is given by

$$f_{X_1,\,X_2}(x_1,\, x_2) = \begin{cases} \frac{e^{-(\lambda_1+\lambda_2)}\lambda_1^{x_1}\lambda_2^{x_2}}{x_1!x_2!} & \text{if } x_1 = 0,\, 1,\, \ldots;\, x_2 = 0,\, 1,\, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $S_{X_1,\,X_2} = \{0,\, 1,\, 2,\, \ldots\} \times \{0,\, 1,\, 2,\, \ldots\}$, which implies that $S_Y = \{0,\, 1,\, 2,\, \ldots\}$. For $y \in S_Y$, $A_y = \{(x,\, y - x) : x = 0,\, 1,\, \ldots,\, y\}$. Hence, using the Theorem 3.14, for $y \in S_Y$,

$$f_Y(y) = \sum_{(x_1,\,x_2) \in A_y} \frac{e^{-(\lambda_1+\lambda_2)}\lambda_1^{x_1}\lambda_2^{x_2}}{x_1!x_2!} = \frac{e^{-(\lambda_1+\lambda_2)}}{y!} \sum_{x=0}^{y} \binom{y}{x} \lambda_1^x \lambda_2^{y-x} = \frac{1}{y!} e^{-(\lambda_1+\lambda_2)} \left(\lambda_1 + \lambda_2\right)^y.$$

Thus, the PMF of $Y = X_1 + X_2$ is

$$f_Y(y) = \begin{cases} \frac{1}{y!} e^{-(\lambda_1+\lambda_2)} \left(\lambda_1 + \lambda_2\right)^y & \text{if } y = 0,\, 1,\, \ldots \\ 0 & \text{otherwise,} \end{cases}$$

which is PMF of a $P(\lambda_1 + \lambda_2)$. Hence, $X_1 + X_2 \sim P\left(\lambda_1 + \lambda_2\right)$. ||

Example 3.10. Let $X_1 \sim Bin(n_1,\, p)$ and $X_2 \sim Bin(n_2,\, p)$. We also assume that X_1 and X_2 are independent. Suppose that we want to find the PMF of $Y = X_1 + X_2$. Note that X_1 and X_2 are the numbers of successes out of n_1 and n_2 independent Bernoulli trials, respectively. In both the cases the probability of success is p. Therefore, Y is the number of successes out of $n_1 + n_2$ Bernoulli trials with success probability p. As X_1 and X_2 are independent, these $n_1 + n_2$ Bernoulli trials can be assumed to be independent. Hence, the distribution of Y must be $Bin(n_1 + n_2,\, p)$. Let us now check if we get the same distribution using the Theorem 3.14. The JPMF of X_1 and X_2 is

$$f_{X_1,\,X_2}(x_1,\, x_2) = \begin{cases} \binom{n_1}{x_1}\binom{n_2}{x_2} p^{x_1+x_2}(1-p)^{n_1+n_2-x_1-x_2} & \text{if } x_1 = 0,\, 1,\, \ldots,\, n_1;\ x_2 = 0,\, 1,\, \ldots,\, n_2 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $S_{X_1,\,X_2} = \{0,\, 1,\, \ldots,\, n_1\} \times \{0,\, 1,\, \ldots,\, n_2\}$. Without loss of generality, we assume that $n_1 \le n_2$. If not, exchange the roles of X_1 and X_2. Now, $S_Y = \{0,\, 1,\, \ldots,\, n_1 + n_2\}$. For $y \in S_Y$,

$$\begin{aligned} A_y &= \{(x_1,\, x_2) \in S_{X_1,\,X_2} : x_1 + x_2 = y\} \\ &= \begin{cases} \{(x,\, y-x) : x = 0,\, 1,\, \ldots,\, y\} & \text{if } 0 \le y \le n_1 \\ \{(x,\, y-x) : x = 0,\, 1,\, \ldots,\, n_1\} & \text{if } n_1 < y \le n_2 \\ \{(x,\, y-x) : x = y - n_2,\, \ldots,\, n_1\} & \text{if } n_2 < y \le n_1 + n_2. \end{cases} \end{aligned}$$

Hence, for $y \in S_Y$ and $y \le n_1$,

$$f_Y(y) = \sum_{x=0}^{y} \binom{n_1}{x}\binom{n_2}{y-x} p^y (1-p)^{n_1+n_2-y} = \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y}.$$

The last equality can be proved by collecting the coefficient of x^y from both sides of the following expression:

$$(1+x)^{n_1}(1+x)^{n_2} = \left\{\sum_{i=0}^{n_1} \binom{n_1}{i} x^i\right\} \times \left\{\sum_{i=0}^{n_2} \binom{n_2}{i} x^i\right\}.$$

For $y \in S_Y$ and $n_1 < y \le n_2$,

$$f_Y(y) = \sum_{x=0}^{n_1} \binom{n_1}{x}\binom{n_2}{y-x} p^y (1-p)^{n_1+n_2-y} = \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y}.$$

For $y \in S_Y$ and $n_2 < y \le n_1 + n_2$,

$$f_Y(y) = \sum_{x=y-n_2}^{n_1} \binom{n_1}{x}\binom{n_2}{y-x} p^y (1-p)^{n_1+n_2-y} = \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y}.$$

Thus, $X_1 + X_2 \sim Bin\,(n_1 + n_2,\, p)$. Note that independence of X_1 and X_2 and same value of probability of success are important for the result. ||

Theorem 3.15. *Let $\boldsymbol{X} = (X_1,\, \ldots,\, X_n)$ be a continuous random vector with JPDF $f_{\boldsymbol{X}}$.*

1. *Let $y_i = g_i(\boldsymbol{x})$, $i = 1,\, 2,\, \ldots,\, n$ be $\mathbb{R}^n \to \mathbb{R}$ functions such that*

$$\boldsymbol{y} = g(\boldsymbol{x}) = (g_1(\boldsymbol{x}),\, \ldots,\, g_n(\boldsymbol{x}))$$

is one-to-one. That means that there exists the inverse transformation $x_i = h_i(\boldsymbol{y})$, $i = 1,\, 2,\, \ldots,\, n$ defined on the range of the transformation.

2. *Assume that both the mapping and its' inverse are continuous.*

3. *Assume that partial derivatives $\frac{\partial x_i}{\partial y_j}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$, exist and are continuous.*

4. *Assume that the Jacobian of the inverse transformation*

$$J \doteq \det\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j=1,\,2,\,\ldots,\,n} \neq 0$$

on the range of the transformation.

Then $\boldsymbol{Y} = (g_1(\boldsymbol{X}), \ldots, g_n(\boldsymbol{X}))$ is a continuous random vector with JPDF

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(h_1(\boldsymbol{y}), \ldots, h_n(\boldsymbol{y}))|J|.$$

Proof: The proof of this theorem can be done using transformation of variable technique for multiple integration. However, the proof is skipped here. □

Remark 3.7. Note that g is a vector valued function. As g should be one-to-one, the dimension of g should be same as dimension of the argument of g. Though we have written that $g_i : \mathbb{R}^n \to \mathbb{R}$ in the previous theorem, the conclusion of the theorem is valid if we replace $g_i : \mathbb{R}^n \to \mathbb{R}$ by $g_i : S_{\boldsymbol{X}} \to \mathbb{R}$. Moreover, the theorem gives us sufficient conditions for $g(\boldsymbol{X})$ to be a continuous random vector, when $\boldsymbol{X}$ is continuous random vector. Thus, $g(\boldsymbol{X})$ can be a continuous random vector even if the conditions of the previous theorem do not hold true. †

Example 3.11. Let X_1 and X_2 be *i.i.d.* $U(0,\,1)$ random variables. We want to find the JPDF of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Clearly,

$$g_1(x_1,\,x_2) = x_1 + x_2 \quad \text{and} \quad g_2(x_1,\,x_2) = x_1 - x_2.$$

Thus, $\boldsymbol{y} = (y_1,\,y_2) = g(x_1,\,x_2) = (g_1(x_1,\,x_2),\,g_2(x_1,\,x_2)) = (x_1 + x_2,\,x_1 - x_2)$. Now, if $(x_1,\,x_2) \neq (\tilde{x}_1,\,\tilde{x}_2)$, then $g\,(x_1,\,x_2) \neq g(\tilde{x}_1,\,\tilde{x}_2)$. If not, then $x_1 + x_2 = \tilde{x}_1 + \tilde{x}_2$ and $x_1 - x_2 = \tilde{x}_1 - \tilde{x}_2$, which implies $x_1 = \tilde{x}_1$ and $x_2 = \tilde{x}_2$. This is a contradiction. Hence, the function $g(\cdot,\,\cdot)$ is one-to-one. The inverse function is given by $h(y_1,\,y_2) = (h_1(y_1,\,y_2),\,h_2(y_1,\,y_2))$, where $x_1 = h_1(y_1,\,y_2) = \frac{1}{2}(y_1 + y_2)$ and $x_2 = h_2(y_1,\,y_2) = \frac{1}{2}(y_1 - y_2)$. Clearly, both the mapping and inverse mapping are continuous. Now,

$$\frac{\partial x_1}{\partial y_1} = \frac{1}{2},\ \frac{\partial x_1}{\partial y_2} = \frac{1}{2},\ \frac{\partial x_2}{\partial y_1} = \frac{1}{2}, \quad \text{and} \quad \frac{\partial x_2}{\partial y_2} = -\frac{1}{2}.$$

All the partial derivatives are continuous. The Jacobian is

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2} \neq 0.$$

Thus, all the four conditions of the Theorem 3.15 hold, and hence, $\boldsymbol{Y} = (Y_1,\,Y_2)$ is a continuous random vector with JPDF

$$f_{Y_1,\,Y_2}(y_1,\,y_2) = f_{X_1,\,X_2}\left(\frac{1}{2}(y_1 + y_2),\,\frac{1}{2}(y_1 - y_2)\right)\left|-\frac{1}{2}\right|$$

$$= \begin{cases} \frac{1}{2} & \text{if } 0 < y_1 + y_2 < 2,\, 0 < y_1 - y_2 < 2 \\ 0 & \text{otherwise.} \end{cases}$$

Note that in Example 3.7, we have found the distribution of $X_1 + X_2$. You may find the marginal distribution of X_1+X_2 from JPDF above and check if you are getting same marginal distribution. ||

Example 3.12. Let X_1 and X_2 be *i.i.d.* $N(0,\,1)$ random variables. We want to find the PDF of $Y_1 = X_1/X_2$. Note that we cannot use Theorem 3.15 directly here as we have a single function $g_1(x_1,\,x_2) = \frac{x_1}{x_2}$. Thus, we need to bring an auxiliary new function $g_2(x_1,\,x_2)$ such that $g(x_1,\,x_2) = (g_1(x_1,\,x_2),\, g_2(x_1,\,x_2))$ satisfies all the conditions of Theorem 3.15. Let us take $g_2(x_1,\,x_2) = x_2$. Clearly, $g(x_1,\,x_2)$ is a one-to-one function. Here, the inverse function is $h(y_1,\,y_2) = (h_1(y_1,\,y_2),\, h_2(y_1,\,y_2))$, where $x_1 = h_1(y_1,\,y_2) = y_1y_2$ and $x_2 = h_2(y_1,\,y_2) = y_2$. It is easy to see that mapping g and its' inverse are continuous. Also,

$$\frac{\partial x_1}{\partial y_1} = y_2,\ \frac{\partial x_1}{\partial y_2} = y_1,\ \frac{\partial x_2}{\partial y_1} = 0, \quad \text{and} \quad \frac{\partial x_2}{\partial y_2} = 1.$$

All the partial derivatives are continuous. Hence, the Jacobian is

$$J = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2.$$

Thus, all the four conditions of the Theorem 3.15 hold, and hence, $\boldsymbol{Y} = \left(\frac{X_1}{X_2},\, X_2\right)$ is a continuous random vector with JPDF

$$f_{Y_1,Y_2}(y_1,\,y_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(1+y_1^2)y_2^2}\,|y_2| \quad \text{for } (y_1,\,y_2) \in \mathbb{R}^2.$$

Now, we can find the marginal PDF of Y_1 from the JPDF of $(Y_1,\,Y_2)$. The marginal PDF of Y_1 is given by

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} \frac{|y_2|}{2\pi} e^{-\frac{1}{2}(1+y_1^2)y_2^2} dy_2 = \frac{1}{\pi}\int_0^{\infty} y_2 e^{-\frac{1}{2}(1+y_1^2)y_2^2} dy_2 = \frac{1}{\pi(1+y_1^2)}$$

for all $y_1 \in \mathbb{R}$. Thus, $Y_1 \sim \textit{Cauchy}(0,\,1)$. ||

Theorem 3.16. *If X and Y are independent, then $g(X)$ and $h(Y)$ are also independent.*

Proof: The exact proof of the theorem cannot be given in the course. However, intuitively it makes sense. X and Y are independent. That means that there is no effect of one of the RVs on the other. Now, g being a function of X only and h being a function of Y only, there should be no effect of $g(X)$ on $h(Y)$ and vice versa. □

3.8.3 Technique 3

The Technique 3 depends on the MGF. Hence, first we need to define the MGF of a random vector.

Definition 3.13 (Moment Generating Function)**.** *Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a random vector. The MGF of $\boldsymbol{X}$ at $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$ is defined by*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = E\left(\exp\left(\sum_{i=1}^{n} t_iX_i\right)\right)$$

provided the expectation exists in a neighborhood of origin $\mathbf{0} = (0,\,0,\,\ldots,\,0)$.

Theorem 3.17. $E(X_1^{r_1}X_2^{r_2}\cdots X_n^{r_n}) = \frac{\partial^{r_1+r_2+\ldots+r_n}}{\partial t_1^{r_1}\partial t_2^{r_2}\ldots\partial t_n^{r_n}} M_{\boldsymbol{X}}(\boldsymbol{t})\Big|_{\boldsymbol{t}=0}$.

Proof: The proof is out of scope of the course. □

Theorem 3.18. *X and Y are independent iff $M_{X,Y}(t_1, t_2) = M_X(t_1)M_Y(t_2)$ in a neighborhood of the origin.*

Proof: The proof is out of scope of the course. □

Note that if X and Y are independent, then using Theorems 3.11, it is straight forward to see that $M_{X,Y}(t,\, s) = M_X(t)M_Y(s)$. Also, note that $E\left(g(X)h(Y)\right) = E\left(g(X)\right)E\left(h(Y)\right)$ for some functions g and h does not imply that X and Y are independent. In particular $E(XY) = E(X)E(Y)$ does not imply X and Y are independent. Please revisit Example 3.6 in this regard.

Definition 3.14. *Two n-dimensional random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to have the same distribution, denoted by $\boldsymbol{X} \overset{d}{=} \boldsymbol{Y}$, if $F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{\boldsymbol{Y}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

Theorem 3.19. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two n-dimensional random vectors. Let $M_{\boldsymbol{X}}(\boldsymbol{t}) = M_{\boldsymbol{Y}}(\boldsymbol{t})$ for all $\boldsymbol{t}$ in a neighborhood around $\boldsymbol{0}$, then $\boldsymbol{X} \overset{d}{=} \boldsymbol{Y}$.*

Proof: The proof is out of scope of the course. □

Example 3.13. Let X_i, $i = 1, 2, \ldots, k$ be independent $Bin(n_i, p)$ RVs. Let us try to find the distribution of $Y = \sum_{i=1}^{k} X_i$. Now, the MGF of Y is

$$M_Y(t) = E\left(e^{tY}\right) = E\left(\exp\left(t\sum_{i=1}^{k} X_i\right)\right) = E\left(\prod_{i=1}^{k} e^{tX_i}\right) = \prod_{i=1}^{k} E(e^{tX_i}) = \prod_{i=1}^{k} M_{X_i}(t).$$

The fourth equality is true as the RVs $X_1,\, X_2,\, \ldots,\, X_k$ are independent. In Example 2.40, we have seen that the MGF of $X \sim Bin(n,\, p)$ is $M_X(t) = (1-p+pe^t)^n$ for all $t \in \mathbb{R}$. Thus, the MGF of Y is

$$M_Y(t) = \prod_{i=1}^{k} (1-p+pe^t)^{n_i} = (1-p+pe^t)^{\sum_{i=1}^{k} n_i}$$

for $t \in \mathbb{R}$. Let $Z \sim Bin\left(\sum_{i=1}^{k} n_i,\, p\right)$, then $M_Z(t) = M_Y(t)$ for all $t \in \mathbb{R}$. Thus, $Y \overset{d}{=} Z \sim Bin\left(\sum_{i=1}^{k} n_i,\, p\right)$. Note that this example is an extension of Example 3.10. ||

Example 3.14. Let $X_1,\, X_2,\, \ldots,\, X_k \overset{i.i.d.}{\sim} Exp(\lambda)$ and $Y = \sum_{i=1}^{k} X_i$. Then the MGF of Y is

$$M_Y(t) = \prod_{i=1}^{k} M_{X_i}(t) = \left[M_{X_1}(t)\right]^k = \left(1 - \frac{t}{\lambda}\right)^{-k}$$

for all $t < \lambda$. The second equality is due to the fact that X_i has same distribution for all $i = 1,\, 2,\, \ldots,\, k$. The third equality is hold true form Example 2.41. Let $Z \sim Gamma(k, \lambda)$. Then $M_Z(t) = M_Y(t)$ for all $t < \lambda$. Hence, $Y \sim Gamma(k,\, \lambda)$. ||

Example 3.15. Let X_i, $i = 1, 2, \ldots, k$ be independent $N(\mu_i, \sigma_i^2)$ RVs. Then $\sum_{i=1}^k X_i \sim N\left(\sum_{i=1}^k \mu_i, \sum_{i=1}^k \sigma_i^2\right)$. This can be proved following the same technique as the last example. I am leaving it as an exercise. ||

Definition 3.15 (Expectation of a Random Vector)**.** *Expectation of a random vector is given by*

$$E(\boldsymbol{X}) = (EX_1, EX_2, \ldots, EX_n)' = \boldsymbol{\mu}.$$

Definition 3.16 (Variance-Covariance Matrix of a Random Vector)**.** *The variance-covariance matrix of a n-dimensional random vector, denoted by* Σ, *is defined by*

$$\Sigma = [Cov(X_i, X_j)]_{i,j=1}^n = E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'.$$

3.9 Conditional Distribution

3.9.1 For Discrete Random Vector

Definition 3.17. *Let (X, Y) be a discrete random vector with JPMF $f_{X,Y}(\cdot, \cdot)$. Suppose the marginal PMF of Y is $f_Y(\cdot)$. The conditional PMF of X, given $Y = y$ is defined by*

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

provided $f_Y(y) > 0$.

Note that $f_{X,Y}(x, y) = P(X = x, Y = y)$ and $f_Y(y) = P(Y = y)$. Thus, the conditional PMF of X given $Y = y$ is $P(X = x|Y = y)$. As we know that $P(A|B)$ is defined if $P(B) > 0$, here we need the condition that $f_Y(y) = P(Y = y) > 0$. Hence, $f_{X|Y}(x|y)$ is only defined for $y \in S_Y$.

Example 3.16. Let $X_1 \sim Poi(\lambda_1)$, $X_2 \sim Poi(\lambda_2)$. Also, assume that X_1 and X_2 are independent. In Example 3.9, we have seen that $X_1 + X_2 \sim Poi(\lambda_1 + \lambda_2)$ and the support of $X_1 + X_2$ is $S = \{0, 1, 2, \ldots\}$. Hence, the conditional PMF of X_1 given $X_1 + X_2 = y$ is defined for all $y \in S$, and the conditional PMF of X_1 given $X_1 + X_2 = y$ is given by

$$\begin{aligned}
f_{X_1|X_1+X_2}(x|y) &= \frac{f_{X_1, X_1+X_2}(x, y)}{f_{X_1+X_2}(y)} \\
&= \frac{P(X_1 = x, X_1 + X_2 = y)}{P(X_1 + X_2 = y)} \\
&= \frac{P(X_1 = x, X_2 = y - x)}{P(X_1 + X_2 = y)} \\
&= \frac{P(X_1 = x)\, P(X_2 = y - x)}{P(X_1 + X_2 = y)}, \quad \text{as } X_1 \text{ and } X_2 \text{ are independent} \\
&= \begin{cases} \frac{e^{-\lambda_1} \frac{\lambda_1^x}{x!} \times e^{-\lambda_2} \frac{\lambda_2^{y-x}}{(y-x)!}}{e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^y}{y!}} & \text{if } x = 0, 1, 2 \ldots, y \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \binom{y}{x} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x \left(1 - \frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{y-x} & \text{if } x = 0, 1, 2 \ldots, y \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Thus, $X_1|X_1+X_2 = y \sim Bin\left(y,\, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$. Note that the support of the conditional PMF is $\{0,\, 1,\, \ldots,\, y\}$. $||$

Definition 3.18. *The conditional CDF of X given $Y = y$ is defined by*

$$F_{X|Y}(x|y) = P(X \leq x|Y = y) = \sum_{\{u \leq x : (u,y) \in S_{X,Y}\}} f_{X|Y}(u|y),$$

provided $f_Y(y) > 0$.

Definition 3.19 (Conditional Expectation for Discrete Random Vector)**.** *The conditional expectation of $h(X)$ given $Y = y$ is defined by*

$$E(h(X)|Y = y) = \sum_{\{x : (x,y) \in S_{X,Y}\}} h(x) f_{X|Y}(x|y),$$

provided it is absolutely summable.

Remark 3.8. Note that for fixed $y \in S_Y$, $f_{X|Y}(\cdot|y)$ is PMF. Thus, conditional expectation is an expectation with respect to the distribution specified by the PMF $f_{X|Y}(\cdot|y)$, and hence, conditional expectation satisfies all the properties of expectation. For example, if $h_1(x) \leq h_2(x)$ for all $x \in \mathbb{R}$, then

$$E\left(h_1(X)|Y = y\right) \leq E\left(h_2(X)|Y = y\right),$$

provided the expectations exist. †

Example 3.17. Let $X \sim Poi(\lambda_1)$, $Y \sim Poi(\lambda_2)$. Let X and Y be independent. In Example 3.16, we have seen that $X|X+Y = n \sim Bin(n,\, \frac{\lambda_1}{\lambda_1+\lambda_2})$ for all $n = 0,\, 1,\, \ldots$. Hence, the conditional expectation of X given $X+Y = n$ is $\frac{n\lambda_1}{\lambda_1+\lambda_2}$. $||$

Example 3.18. Suppose that a system has n components. Suppose that on a rainy day, component i functions with probability p_i, $i = 1,\, 2,\, \ldots,\, n$. Also, assume that the components work independently. We want to calculate the conditional expected number of components that will function tomorrow given that it will rain tomorrow. Again we will use the indicator RVs as we used in Example 3.4 to count the number of components that will work tomorrow. Let

$$X_i = \begin{cases} 1 & \text{if component } i \text{ functions tomorrow} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad Y = \begin{cases} 1 & \text{if it rains tomorrow} \\ 0 & \text{otherwise.} \end{cases}$$

Then the desired expectation can be obtained as follows:

$$E\left[\sum_{i=1}^{n} X_i|Y = 1\right] = \sum_{i=1}^{n} E\left(X_i|Y = 1\right) = \sum_{i=1}^{n} p_i.$$

The last equality is due to the fact that $P\left(X_i|Y = 1\right) = p_i$ for all $i = 1,\, 2,\, \ldots,\, n$. $||$

Theorem 3.20. *If X and Y are independent DRVs, then $f_{X|Y}(x|y) = f_X(x)$ for all $x \in \mathbb{R}$ and $y \in S_Y$.*

Proof: The proof is straight forward from the definition of conditional PMF. □

3.9.2 For Continuous Random Vector

Let $(X,\, Y)$ be a continuous random vector. Note that, in this case, Y is a CRV, and hence, $P(Y = y) = 0$ for all $y \in \mathbb{R}$. As a result, we cannot define the conditional probabilities in the same way as we defined for the discrete random vector in the previous subsection. Like CRV or continuous random vector, we will first define the conditional CDF for a continuous random vector $(X,\, Y)$, then conditional PDF.

Definition 3.20 (Conditional CDF)**.** *Let $(X,\, Y)$ be a continuous random vector. The conditional CDF of X given $Y = y$ is defined as*

$$F_{X|Y}(x|y) = \lim_{\epsilon \downarrow 0} P(X \leq x | Y \in (y - \epsilon, y + \epsilon]),$$

provided the limit exists.

Note that CDF of a random variable, X, is defined by $P\left(X \leq x\right)$. Ideally, we want to see what is the value of $P\left(X \leq x|Y = y\right)$. However, when $(X,\, Y)$ is continuous random vector, we have a problem to define $P\left(X \leq x|Y = y\right)$ as $P\left(Y = y\right) = 0$ for all $y \in \mathbb{R}$. One of the way to overcome this difficulty is to proceed as follows. We can replace the event $\{Y = y\}$ by $Y \in (y - \epsilon,\, y + \epsilon]$ and then take limit that ϵ drops to zero. Of course, this makes sense if the limit exists. Motivated by this intuition, we have the previous definition of conditional CDF of X given $Y = y$ for a continuous random vector $(X,\, Y)$.

Definition 3.21 (Conditional PDF)**.** *Let $(X,\, Y)$ be a continuous random vector with conditional CDF $F_{X|Y}(\cdot|y)$ of X given $Y = y$. Define the conditional PDF of X given $Y = y$, $f_{X|Y}(x|y)$, as the non-negative integrable function satisfying*

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(t|y)dt \quad \text{for all } x \in \mathbb{R}.$$

Theorem 3.21. *Let $f_{X,Y}$ be the JPDF of (X, Y) and let f_Y be the marginal PDF of Y. If $f_Y(y) > 0$, then the conditional PDF of X given $Y = y$ exists and is given by*

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

Proof: This is not an exact proof, but a overview of it.

$$\begin{aligned}
\lim_{\varepsilon \downarrow 0} P\left(X \leq x | y - \varepsilon < Y \leq y + \varepsilon\right) &= \lim_{\varepsilon \downarrow 0} \frac{P\left(X \leq x,\, y - \varepsilon < Y \leq y + \varepsilon\right)}{P\left(y - \varepsilon < Y \leq y + \varepsilon\right)} \\
&= \lim_{\varepsilon \downarrow 0} \frac{\int_{-\infty}^{x} \int_{y-\varepsilon}^{y+\varepsilon} f_{X,Y}(t,\, s)dsdt}{\int_{y-\varepsilon}^{y+\varepsilon} f_Y(s)ds} \\
&= \int_{-\infty}^{x} \frac{\lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_{y-\varepsilon}^{y+\varepsilon} f_{X,Y}(t,\, s)ds}{\lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_{y-\varepsilon}^{y+\varepsilon} f_Y(s)ds} dt \\
&= \int_{-\infty}^{x} \frac{f_{X,Y}(t,\, y)}{f_Y(y)} dt.
\end{aligned}$$

The last equality is due to fundamental theorems of calculus. Thus, the conditional PDF is given by $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ for those values of $y \in \mathbb{R}$ for which $f_Y(y) > 0$. □

Definition 3.22 (Conditional Expectation for Continuous Random Vector)**.** *The conditional expectation of $h(X)$ given $Y = y$, is defined for all values of y such that $f_Y(y) > 0$, by*

$$E(h(X)|Y = y) = \int_{-\infty}^{\infty} h(x) f_{X|Y}(x|y) dx,$$

provided it is absolutely integrable.

Remark 3.9. Note that for fixed $y \in S_Y$, $f_{X|Y}(\cdot|y)$ is a PDF. Therefore, conditional expectation is an expectation with respect to the PDF $f_{X|Y}(\cdot|y)$. Thus, $E(X|Y = y)$ satisfies all properties of unconditional expectation. †

Example 3.19. Suppose the JPDF of (X, Y) is given by

$$f_{X,Y}(x, y) = \begin{cases} 6xy(2 - x - y) & 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

The marginal PDF of Y is

$$f_Y(y) = \begin{cases} \frac{1}{6} y(4 - 3y) & \text{if } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the conditional PDF of X given $Y = y \in (0,\, 1)$ is given by

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,\, y)}{f_Y(y)} \\ &= \begin{cases} \frac{6x(2-x-y)}{4-3y} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The conditional expectation of X given $Y = y \in (0,\, 1)$ is

$$E\left(X|Y = y\right) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx = \frac{6}{4 - 3y} \int_0^1 x^2(2 - x - y) dx = \frac{5 - 4y}{2\left(4 - 3y\right)}.$$

Note that for conditional PDF, the ranges of both of x and y are important and need to mention unambiguously. Similarly for computing conditional expectation, we need the appropriate ranges of x and y. ||

Example 3.20. Let the joint PDF of $(X,\, Y)$ be

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{2} y e^{-xy} & \text{if } 0 < x < \infty,\, 0 < y < 2 \\ 0 & \text{otherwise.} \end{cases}$$

The marginal PDF of Y is

$$f_Y(y) = \begin{cases} \frac{1}{2} & \text{if } 0 < y < 2 \\ 0 & \text{otherwise.} \end{cases}$$

For $y \in (0,\, 2)$, the conditional PDF of X given $Y = y$ is

$$f_{X|Y}(x|y) = \begin{cases} y e^{-yx} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$E\left(e^{\frac{X}{2}}|Y=1\right)=\int_0^\infty e^{-x+\frac{x}{2}}dx=2.$$

Note that, in this example, $Y \sim U(0,\, 2)$ and $X|Y=y \sim Exp(y)$ for $y \in (0,\, 2)$. ||

Theorem 3.22. *If $(X,\, Y)$ is a continuous random vector such that X and Y are independent random variables, then $f_{X|Y}(x|y)=f_X(x)$ for all $x \in \mathbb{R}$ and for all $y \in S_Y$.*

Proof: The proof is straight forward from the Theorem 3.21. □

3.9.3 Computing Expectation by Conditioning

Suppose that (X, Y) is either a discrete random vector or a continuous random vector. Then the conditional expectation $E(X|Y=y)$ is a function of y. Let we denote $g(y)=E(X|Y=y)$. Then $g(Y)$ is a function of RV Y. Thus, $g(Y)=E(X|Y)$ is again a random variable. With this understanding, we have the following theorem.

Theorem 3.23. *$E(X)=E(E(X|Y))$.*

Proof: We will prove it for continuous random vector $(X,\, Y)$. For the discrete random vector, the proof can be obtained by replacing the integration sign by summation sign.

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

□

In the previous theorem, the outside expectation is with respect to the distribution of Y, as $E\left(X|Y\right)$ is a function of Y. This theorem can be used to solve many problems. This theorem states that we can compute the average of different parts of a population separately and then take an weighted sum of those average to obtain the overall average. Let there be three columns of seating arrangement (like some of the lectures halls at IITG) in a class and we want to find the average of heights of the students in the class. Let $\overline{x}_i$ denote the average of the height of the students seating in the ith column and n_i be the number of students who is seating in ith column. Then the overall average is

$$\overline{x}=\frac{n_1}{n}\overline{x}_1+\frac{n_2}{n}\overline{x}_2+\frac{n_3}{n}\overline{x}_3,$$

where $n=n_1+n_2+n_3$. Note that $\frac{n_i}{n}$ can be interpreted as the probability that a student in column i and $\overline{x}_i$ is the conditional expectation of height given that the student is in the column i. Therefore, the overall average is $EE\left(X|Y\right)$, where X denotes the height of a student and Y is an indicator of the column number. We can take $Y=i$ if the student is in the ith column. Thus, the above theorem is a generalization of what we have learned in school.

Though we have discussed expectations when (X, Y) is either discrete random vector or continuous random vector, the above theorem is still valid if one of them is DRV and other is CRV. We will see some applications where one of them is CRV and other one is DRV. Of course, we will not go for proper definition (which is out of scope of this course) or proof in these cases. If Y is a DRV, then

$$E(X) = EE(X|Y) = \sum_{y \in S_Y} E(X|Y = y)\, f_Y(y).$$

If Y is a CRV, then

$$E(X) = EE(X|Y) = \int_{-\infty}^{\infty} E(X|Y = y)\, f_Y(y)dy.$$

Example 3.21. Virat will read either one chapter of his probability book or one chapter of his history book. If the number of misprints in a chapter of his probability and history book is Poisson with mean 2 and 5 respectively, then assuming that Virat is equally likely to choose either book, we can compute the expected number of misprints that he will come across using the above theorem. Let X denote the number of misprint and

$$Y = \begin{cases} 1 & \text{if Virat read the probability book} \\ 2 & \text{if Virat read the history book.} \end{cases}$$

We need to find $E(X)$. Note that $P(Y = 1) = P(Y = 2) = \frac{1}{2}$, $E(X|Y = 1) = 2$ and $E(X|Y = 2) = 5$. Hence,

$$E(X) = EE(X|Y) = P(Y = 1)\, E(X|Y = 1) + P(Y = 2)\, E(X|Y = 2) = \frac{1}{2}(2 + 5) = 3.5.$$

Thus, expected number of misprint that Virat will come across is 3.5. ||

Theorem 3.24. *$E(X - E(X|Y))^2 \leq E(X - f(Y))^2$ for any function f.*

Proof: Let us denote $\mu(Y) = E(X|Y)$. Then

$$\begin{aligned} E(X - f(Y))^2 &= E(X - \mu(Y) + \mu(Y) - f(Y))^2 \\ &= E(X - \mu(Y))^2 + E(\mu(Y) - f(Y))^2 + 2E[(X - \mu(Y))(\mu(Y) - f(Y))]. \end{aligned}$$

Now,

$$\begin{aligned} E[(X - \mu(Y))(\mu(Y) - f(Y))] &= EE[(X - \mu(Y))(\mu(Y) - f(Y))|Y] \\ &= E[(\mu(Y) - f(Y))E(X - \mu(Y)|Y)] \\ &= E[(\mu(Y) - f(Y))(\mu(Y) - \mu(Y))] \\ &= 0. \end{aligned}$$

The first equality is due to the Theorem 3.23. For the second equality, notice that $\mu(Y) - f(Y)$, being a function of Y only, acts as a constant when Y is given. Hence, $\mu(Y) - f(Y)$ comes out of the conditional expectation. Thus,

$$E(X - f(Y))^2 = E(X - \mu(Y))^2 + E(\mu(Y) - f(Y))^2 \geq E(X - \mu(Y))^2,$$

as $E(\mu(Y) - f(Y))^2 \geq 0$. The equality holds if and only if $E(\mu(Y) - f(Y))^2 = 0$. It can be shown that $E(\mu(Y) - f(Y))^2 = 0$ if and only if $f(Y) = \mu(Y) = E(X|Y)$. □

Recall the Theorem 2.12, which states that if we do not have any extra information, then the "best estimate" of X is $E(X)$. The Theorem 3.24 states that if we have information on the RV Y, then the "best estimate" of X is changes and becomes $E\left(X|Y\right)$.

Definition 3.23 (Conditional Variance)**.** *Let (X, Y) be a random vector. Then the conditional variance of X given $Y = y$ is defined by*

$$Var(X|Y = y) = E((X - E(X|Y))^2|Y = y) = E(X^2|Y = y) - (E(X|Y = y))^2.$$

Like expectation, $Var(X|Y) = \sigma^2(Y)$ is a RV, where $\sigma^2(y) = Var(X|Y = y)$. Then we have the following theorem. The following theorem says that the overall variance can be computed by calculating variances and expectations of different parts and then aggregating.

Theorem 3.25. $Var(X) = E(Var(X|Y)) + Var(E(X|Y))$.

Proof: Let $\mu(Y) = E\left(X|Y\right)$ and $\mu = E\left(X\right) = EE\left(X|Y\right)$. Then

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

Now, using Theorem 3.23,

$$E\left[\left(X - \mu(Y)\right)\left(\mu(Y) - \mu\right)\right] = EE\left[\left(X - \mu(Y)\right)\left(\mu(Y) - \mu\right)|Y\right] = 0.$$

Thus,

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

Note that it is easy to remember the formula. On the right hand side, one is expectation of conditional variance and another is variance of conditional expectation. □

Example 3.22. Let $X_0, X_1, X_2, \ldots$ be a sequence of *i.i.d.* RVs with mean μ and variance σ^2. Let $N \sim Bin(n, p)$ and is independent of X_i's for all $i = 0,\ 1,\ \ldots$. Define $S = \sum_{i=0}^{N} X_i$. Note that the RV S is the sum of random number of RVs. This type of RVs are called compound RVs. Compound random variables are quite important in many practical situations. For example, consider a car insurance company. The number of accidents that a customer meets in a year is a RV. Let N denotes the number of accident in a year. Now, assume that X_i denotes the claim by the customer after the ith accident. Then S is the total claim made by the customer. Now, it is important for the insurance company to have an idea of average and variance of claims made by a customer. Let us try to compute $E(S)$ and $Var(S)$. Note that

$$E\left(S|N = n\right) = E\left(\sum_{i=0}^{N} X_i|N = n\right) = E\left(\sum_{i=0}^{n} X_i|N = n\right) = E\left(\sum_{i=0}^{n} X_i\right) = (n+1)\mu.$$

The second equality is true as under the condition $N = n$, the sum has n components. The third equality is true due to the fact that N and X_i's are independent. Note that $\sum_{i=1}^{N} X_i$

and N are not independent. However, when we put a specific value of N in $\sum_{i=1}^{N} X_i$ to get $\sum_{i=1}^{n} X_i$, then the later does not involve N and becomes independent of N. Thus, we have $E\left(S|N\right) = (N+1)\mu$. Hence,

$$E(S) = EE(S|N) = E\left[(N+1)\mu\right] = (np+1)\mu.$$

Now,

$$Var(S|N=n) = Var\left(\sum_{i=0}^{N} X_i|N=n\right) = Var\left(\sum_{i=0}^{n} X_i|N=n\right) = Var\left(\sum_{i=0}^{n} X_i\right) = (n{+}1)\sigma^2.$$

Thus, $Var(S|N) = (N+1)\sigma^2$. Hence,

$$\begin{aligned}Var(S) &= E\left(Var(S|N)\right) + Var\left(E(S|N)\right)\\ &= E\left[(N+1)\sigma^2\right] + Var\left[(N+1)\mu\right]\\ &= (np+1)\sigma^2 + np(1-p)\mu^2.\end{aligned}$$

||

Theorem 3.23 can be used to compute probability by conditioning. We have seen that $P\left(X \in A\right) = E\left(I_A(X)\right)$, where I_A is the indicator function of the set A. Also, note that $E\left(I_A(X)|Y=y\right) = P\left(X \in A|Y=y\right)$. Therefore, we can write

$$\begin{aligned}P(A) = P\left(X \in A\right) &= E\left(I_A(X)\right) = EE\left(I_A(X)|Y\right)\\ &= \begin{cases}\displaystyle\sum_{y\in S_Y} P(A|Y=y)P(Y=y) & \text{for } Y \text{ discrete}\\ \displaystyle\int_{-\infty}^{\infty} P(A|Y=y)f_Y(y)dy & \text{for } Y \text{ continuous.}\end{cases}\end{aligned}$$

When Y is a DRV, $P\left(E\right) = \sum_{y\in S_Y} P\left(E|Y=y\right)P(Y=y)$ can be concluded from the Theorem 1.17. Of course, the result for CRV cannot be obtained form the Theorem 1.17.

Example 3.23. Let X and Y be independent CRVs having PDFs f_X and f_Y, respectively. Then

$$\begin{aligned}P\left(X<Y\right) &= \int_{-\infty}^{\infty} P\left(X<Y|Y=y\right)f_Y(y)dy\\ &= \int_{-\infty}^{\infty} P\left(X<y|Y=y\right)f_Y(y)dy\\ &= \int_{-\infty}^{\infty} P\left(X<y\right)f_Y(y)dy\\ &= \int_{-\infty}^{\infty} F_X(y)f_Y(y)dy,\end{aligned}$$

where $F_X(\cdot)$ is the CDF corresponding to $f_X(\cdot)$. ||

Example 3.24. Let X and Y be *i.i.d.* CRVs having common PDF $f(\cdot)$ and CDF $F(\cdot)$. Then using the last example

$$P\left(X<Y\right) = \int_{-\infty}^{\infty} F(y)f(y)dy = \frac{1}{2}.$$

Now, as X and Y are *i.i.d.*, $P(Y<X) = P(X>Y) = \frac{1}{2}$. Thus, $P(X=Y) = 1 - P\left(X<Y\right) - P\left(X>Y\right) = 0$. ||

Example 3.25. Suppose X and Y are two independent RVs, either discrete or continuous. Let us study the RV $Z = X + Y$ and try to see if this is a CRV or DRV

We know that (X, Y) is a discrete random vector if X and Y are DRVs, and hence, $Z = X + Y$ is a DRV. The PMF of Z, for $z \in \mathbb{R}$, is

$$\begin{aligned} f_Z(z) &= P\left(X + Y = z\right) \\ &= \sum_{y \in S_y} P\left(X + Y = z | Y = y\right) P(Y = y) \\ &= \sum_{y \in S_Y} P\left(X + y = z | Y = y\right) f_Y(y) \\ &= \sum_{y \in S_Y} P\left(X = z - y\right) f_Y(y) \\ &= \sum_{y \in S_Y} f_X(z - y) f_Y(y). \end{aligned}$$

Now, assume that X and Y are CRVs. Let us first find the CDF of Z and then check what type of RV Z is. The CDF of Z, for $z \in \mathbb{R}$, is

$$\begin{aligned} F_Z(z) &= P\left(X + Y \leq z\right) \\ &= \int_{-\infty}^{\infty} P\left(X + Y \leq z | Y = y\right) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} P\left(X + y \leq z | Y = y\right) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} P\left(X \leq z - y\right) f_Y(y) dy, \quad \text{as } X \text{ and } Y \text{ are independent} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z} f_X(x' - y) f_Y(y) dx' dy, \quad \text{taking } x' = x + y \\ &= \int_{-\infty}^{z} \left\{ \int_{-\infty}^{\infty} f_X(x' - y) f_Y(y) dy \right\} dx' \quad \text{for all } z \in \mathbb{R}. \end{aligned}$$

Thus, $X + Y$ is a CRV with PDF

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy \quad \text{for all } z \in \mathbb{R}.$$

Note that changing the role of X and Y, we can write the PDF of Z as

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_Y(z - x) f_X(x) dx \quad \text{for all } z \in \mathbb{R}.$$

Now, assume that X is a CRV and Y is a DRV. Then the CDF of $Z = X + Y$ is

$$\begin{aligned} F_Z(z) &= P\left(X + Y \leq z\right) \\ &= \sum_{y \in S_Y} P\left(X + Y \leq z | Y = y\right) f_Y(y) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{y \in S_Y} \int_{-\infty}^{z-y} f_X(x) f_Y(y) dx \\
&= \sum_{y \in S_Y} \int_{-\infty}^{z} f_X(x'-y) f_Y(y) dx', \quad \text{taking } x' = x+y \\
&= \int_{-\infty}^{z} \left\{ \sum_{y \in S_Y} f_X(x'-y) f_Y(y) \right\} dx' \quad \text{for all } z \in \mathbb{R}.
\end{aligned}
$$

Therefore, $X+Y$ is a CRV with PDF

$$f_{X+Y}(z) = \sum_{y \in S_Y} f_X(z-y) f_Y(y) \quad \text{for all } z \in \mathbb{R}.$$

To summarize, if X and Y are independent, then the RV $X+Y$ is continuous if at least one of X or Y is CRV. If both of them are DRVs, then $X+Y$ is a DRV. ||

Definition 3.24 (Conditional Expectation for given Event)**.** *Let (X, Y) be a random vector. Then*

$$E\left(h(X,\, Y) | (X,\, Y) \in A\right) = \frac{E(h(X,\, Y) I_A(X,\, Y))}{P\left((X,\, Y) \in A)\right)}.$$

Example 3.26. Let $X \sim Exp(1)$. Then

$$E\left(X | X \geq 2\right) = \frac{E\left(X I_{[2,\,\infty)}(X)\right)}{P\left(X \geq 2\right)} = \frac{\displaystyle\int_0^{\infty} x I_{[2,\,\infty)}(x) e^{-x} dx}{\displaystyle\int_2^{\infty} e^{-x} dx} = e^2 \int_2^{\infty} x e^{-x} dx = 3.$$

||

Example 3.27. $(X,\, Y)$ is uniform on unit square. Then

$$E\left(X | X+Y > 1\right) = \frac{E\left(X I_{(1,\,\infty)}(X+Y)\right)}{P\left(X+Y > 1\right)} = \frac{\displaystyle\int_0^1 \int_{1-x}^1 x dy dx}{\displaystyle\int_0^1 \int_{1-y}^1 dx dy} = \frac{2}{3}.$$

||

Example 3.28. A rod of length l is broken into two parts. Then the expected length of the shorter part is

$$E\left(X | X < \frac{l}{2}\right),$$

where $X \sim U(0,\, l)$. Thus, the expected length of the shorter part is

$$E\left(X | X < \frac{l}{2}\right) = \frac{E\left(X I_{(-\infty,\, \frac{l}{2})}(X)\right)}{P\left(X < \frac{l}{2}\right)} = \frac{\displaystyle\frac{1}{l}\int_0^{\frac{l}{2}} x dx}{\displaystyle\frac{1}{l}\int_0^{\frac{l}{2}} dx} = \frac{l}{4}.$$

An alternative formulation is as follows: The required quantity is $E\left[\min\left\{X,\, l-X\right\}\right]$. Please calculate and check if you are getting the same value. ||

Chapter 4

Limit Theorems

This chapter will deal with convergence properties of sequence of RVs. There are several modes of convergence of sequence of RVs. Here, we will discuss four modes of convergence for a sequence of RVs $\{X_n\}$. Then we will see strong law of large numbers and central limit theorem. These are quite useful concepts in probability. They have applications in different other fields including Statistics. In this chapter, we shall not prove most of the theorems. Our main aim will be to understand the theorems and apply them to solve problems. In rest of the chapter, 1_A denotes the indicator function of the set A.

4.1 Modes of Convergence

Definition 4.1 (Almost Sure Convergence)**.** *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. We say that X_n converges almost surely or with probability (w.p.) 1 to a random variable X if*

$$P\left(\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\}\right) = 1.$$

Example 4.1. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$ and P be a uniform probability (for any interval $I \subseteq \mathcal{S}$, $P(I) =$ length of I). Define the sequence of RVs by

$$X_n(\omega) = 1_{[0,\frac{1}{n}]}(\omega) \quad \text{for all } n = 1,\ 2,\ 3,\ \ldots.$$

Then X_n converges almost surely to the zero RV. Here, the zero RV means a RV, say X, defined on the same probability space $(\mathcal{S},\ \mathcal{F},\ P)$ such that $X(\omega) = 0$ for all $\omega \in \mathcal{S}$. To see it, notice that for any fixed $\omega \in (0,\ 1]$, we can find an n_0 such that $\frac{1}{n} < \omega$ for all $n \geq n_0$. Thus, $X_n(\omega) \to 0 = X(\omega)$ as $n \to \infty$. Therefore, $\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\} = (0,\ 1]$ and hence,

$$P\left(\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\}\right) = P\left((0,\ 1]\right) = 1.$$

Thus, $X_n \to 0$ almost surely. $||$

Definition 4.2 (Convergence in Probability)**.** *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. We say that X_n converges in probability to a random variable X if for any $\epsilon > 0$,*

$$P(|X_n - X| > \epsilon) \to 0 \quad \textit{as } n \to \infty.$$

Example 4.2. Let $\mathcal{S} = [0, 1], \mathcal{F} = \mathcal{B}([0, 1])$ and P be a uniform probability. Define the sequence of RVs using $X_n = 1_{[0, \frac{1}{n}]}$. Then X_n converges in probability to the zero random variable. Let X denote the zero RV defined on the same sample space. To see it, notice that for any fixed $\epsilon > 0$, $|X_n - X| > \epsilon$ only on the interval $[0,\ \frac{1}{n}]$. Thus,

$$P\left(|X_n - X| > \epsilon\right) = \frac{1}{n} \implies \lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0.$$

Therefore, $X_n \to X$ in probability. ||

Example 4.3. Let $\mathcal{S} = [0, 1], \mathcal{F} = \mathcal{B}([0, 1])$ and P be a uniform probability. Define the sequence of RVs using $X_n = n1_{[0, \frac{1}{n}]}$. Then X_n converges in probability to the zero random variable. Let X denote the zero RV defined on the same sample space. To see it, notice that for any fixed $\epsilon > 0$, $|X_n - X| > \epsilon$ only on the interval $[0,\ \frac{1}{n}]$. Thus,

$$P\left(|X_n - X| > \epsilon\right) = \frac{1}{n} \implies \lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0.$$

Therefore, $X_n \to X$ in probability. ||

It may seem that convergence almost surely and convergence in probability are equivalent. However, this is not true, as the following example shows.

Example 4.4. Let $\mathcal{S} = [0, 1], \mathcal{F} = \mathcal{B}([0, 1])$ and P be the uniform probability. Define the sequence of RVs by

$$X_{m,n} = 1_{[\frac{m-1}{2^n},\ \frac{m}{2^n}]} \quad \text{for } m = 1,\ 2,\ \ldots,\ 2^n;\ n = 1,\ 2,\ 3,\ \ldots.$$

Note that $X_{1,1} = 1_{[0,1/2]}, X_{2,1} = 1_{[1/2,1]}, X_{1,2} = 1_{[0,1/4]}, X_{2,2} = 1_{[1/4,1/2]}, X_{3,2} = 1_{[1/2,3/4]}, X_{4,2} = 1_{[3/4,1]}$ and so on. This sequence of RVs $\{X_{m,n}\}$ can be visualized as follows (see Figure 4.1). We start with the interval $[0,\ 1]$. First, we divide the interval into two equal parts, $[0,\ \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. The first RV $X_{1,1}$ is 1 on the first part and 0 on the second part. The second random variable $X_{2,1}$ is 1 on the second part and 0 on the first part. Then, we divide the interval into 2^2 equal parts, *viz.*, $[0,\ \frac{1}{2^2}]$, $[\frac{1}{2^2},\ \frac{2}{2^2}]$, $[\frac{2}{2^2},\ \frac{3}{2^2}]$ and $[\frac{3}{2^2},\ 1]$. Now, the third RV $X_{1,2}$ is 1 on the first part $[1,\ \frac{1}{4}]$ and 0 otherwise. The fourth RV $X_{2,2}$ is 1 on the second part $[\frac{1}{4},\ \frac{1}{2}]$ and 0 otherwise. The fifth RV $X_{3,2}$ equals 1 on the third part $[\frac{1}{2},\ \frac{3}{4}]$ and 0 otherwise. Finally, the sixth RV $X_{4,2}$ is 1 on the fourth part $[\frac{3}{4},\ 1]$ and 0 otherwise. Next, we divide the interval $[0,\ 1]$ into 2^3 equal parts and define the next 8 RVs in the similar manner. This procedure continues.

Let us assume that X be a RV defined on the same probability space and $X = 0$. Then, for any $\epsilon > 0$,

$$P\left(|X_{m,n} - X| > \epsilon\right) = \frac{1}{2^n} \implies \lim_{n\to\infty} P\left(|X_{m,n} - X| > \epsilon\right) = 0.$$

Therefore, $X_{m,n} \to X$ in probability. However, for any fixed $\omega \in \mathcal{S}$, there exists a subsequence of the sequence of real numbers $\{X_{m,n}(\omega)\}$ that converges to one and another subsequence that converges to zero. Therefore, $\{X_{m,n}(\omega)\}$ does not converge for all $\omega \in \mathcal{S}$. Thus,

$$P\left(\{\omega \in \mathcal{S} : X_{m,\,n} \text{ converges}\}\right) = P\left(\emptyset\right) = 0.$$

This shows that $X_{m,n}$ do not converge to any RV almost surely. This example shows that a sequence of RVs, which converges in probability, may not converge almost surely. ||

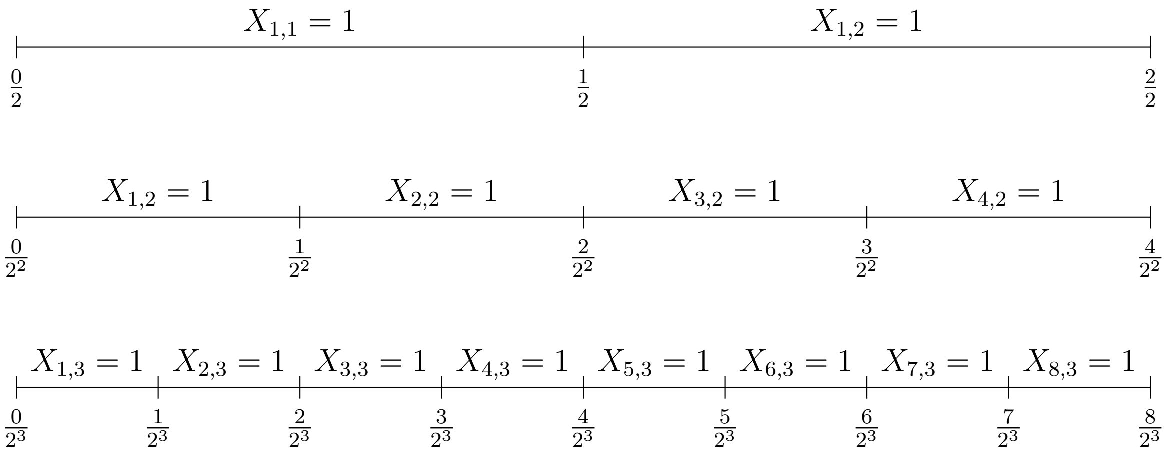


Figure 4.1: Figure for Example 4.4

Definition 4.3 (Convergence in r^{th} Mean)**.** *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. For $r = 1, 2, 3, \ldots$, we say that X_n converges in r^{th} mean to a random variable X if*

$$E|X_n - X|^r \to 0 \quad \text{as } n \to \infty.$$

Example 4.5. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$ and P be a uniform measure. Define $X_n = 1_{[0, \frac{1}{n}]}$. Then X_n converges in 1st mean to the zero random variable. To see it, notice that

$$E|X_n - X| = \frac{1}{n} \to 0 \quad \text{as } n \to \infty,$$

where X is a zero RV defined on the same probability space. $||$

Definition 4.4 (Convergence in Distribution)**.** *Let $\{X_n\}$ be a sequence of RVs and X be a RV. Let $F_n(\cdot)$ and $F(\cdot)$ denote the CDF of X_n and X, respectively. We say that X_n converges in distribution to a random variable X if*

$$F_n(x) \to F(x) \quad \text{as } n \to \infty$$

for all x where F is continuous.

Unlike the first three modes of convergence, here X_n's can be defined on different probability spaces. We are only interested if the sequence of CDFs converges to a CDF. This flexibility makes this mode of convergence very useful.

Example 4.6. Suppose X_ns are random variables such that $P(X_n = \frac{1}{n}) = 1$. Then, the CDF of X_n is

$$F_n(x) = \begin{cases} 0 & \text{if } x < \frac{1}{n} \\ 1 & \text{if } x \geq \frac{1}{n}, \end{cases}$$

which converges pointwise to the function

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

for all $x \neq 0$, which is the point of discontinuity of the function $F(\cdot)$. Now $F(\cdot)$ is the CDF of the RV X, which takes value 0 with probability one. Therefore, X_n converges in distribution to the zero RV. ||

The following theorems states the relation between different modes of convergence.

Theorem 4.1. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. Then $X_n \to X$ in probability if $X_n \to X$ almost surely.*

Proof: This prove is skipped here. □

Theorem 4.2. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. Then $X_n \to X$ in probability if $X_n \to X$ in rth mean for any $r = 1, 2, 3, \ldots$.*

Proof: Let $X_n \to X$ in rth mean. Then, using Markov inequality, for any $\epsilon > 0$,

$$P\left(|X_n - X| > \epsilon\right) \leq \frac{E|X_n - X|^r}{\epsilon^r} \to 0 \quad \text{as } n \to \infty.$$

As probability of an event is always non-negative,

$$P\left(|X_n - X| > \epsilon\right) \to 0 \quad \text{as } n \to \infty.$$

Thus $X_n \to X$ in probability. □

Theorem 4.3. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. Then $X_n \to X$ in distribution if $X_n \to X$ in probability.*

Proof: The proof is skipped here. □

The following figure depicts the relationship between several modes of convergence pictorially. Note that the arrows are one-sided. What about other sides? Moreover, there is no arrows between almost sure convergence and rth mean convergence. The following examples show that in general one mode of convergence does not imply other, whenever there is no directed arrows in the above figure. The Example 4.4 shows that probability convergence does not imply almost sure convergence.

$$\begin{tikzcd} a.s \arrow[dr] & & mean \arrow[dl] \\ & Prob \arrow[d] & \\ & Dist & \end{tikzcd}$$

Example 4.7. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$ and P be a uniform probability. Define the sequence of RVs by

$$X_{m,n} = 1_{[\frac{m-1}{2^n},\, \frac{m}{2^n}]} \quad \text{for } m = 1,\, 2,\, \ldots,\, 2^n;\; n = 1,\, 2,\, 3,\, \ldots.$$

Then

$$E|X_{m,n}| = \frac{1}{2^n} \to 0 \quad \text{as } n \to \infty.$$

Thus, $X_{m,n} \to X = 0$ in 1st mean. However, in Example 4.4, we have seen that $X_{m,n}$ does not convergence almost surely. This example shows that rth mean convergence does not imply almost sure convergence. ||

Example 4.8. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$ and P be a uniform probability. Define $X_n = n1_{[0,\frac{1}{n}]}$. Now, taking $X = 0$,

$$P\left(|X_n - X| > \epsilon\right) = \frac{1}{n} \to 0 \quad \text{as } n \to \infty,$$

for any $\epsilon > 0$. Thus, $X_n \to X$ in probability. Using the logic used in Example 4.1,

$$P\left(\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\}\right) = P((0,\, 1]) = 1.$$

Thus, $X_n \to X$ almost surely. However, X_n does not converge to X in rth mean. To see it, notice that

$$E|X_n - X|^r = n^{r-1} \to \begin{cases} 1 & \text{if } r = 1 \\ \infty & \text{if } r > 1. \end{cases}$$

This example shows that probability convergence or almost sure convergence do not imply rth mean convergence. ||

Example 4.9. Let X be a $N(0,1)$ RV defined on some probability space $(\mathcal{S}, \mathcal{F}, P)$. Define $X_n = X$ for all n. Notice that the CDFs of X_n are same for all $n = 1,\, 2, \ldots$ and is given by $\Phi(\cdot)$. Moreover, the CDFs of X and $-X$ are also $\Phi(\cdot)$. Thus, X_n converges in distribution to $-X$. However, X_n does not converge to $-X$ in probability. To see it, we can proceed as follows: for $\epsilon > 0$,

$$P\left(|X_n + X| \le \epsilon\right) = P\left(2|X| \le \epsilon\right) = 2\Phi\left(\frac{\epsilon}{2}\right) - 1 \neq 1.$$

This example shows that distribution convergence does not imply probability convergence, even if the random variables are defined on the same probability space. ||

Theorem 4.4. *Suppose $\{X_n\}$ is a sequence of RVs defined on a probability space and X_n converges in distribution to some constant c, then X_n also converges in probability to c.*

Proof: As X_n converges to a constant c,

$$F_n(x) \to F(x) = \begin{cases} 0 & \text{if } x < c \\ 1 & \text{if } x \ge c \end{cases}$$

as $n \to \infty$. Now, fix $\varepsilon > 0$. Then,

$$\begin{aligned} 0 \le P\left(|X_n - c| > \varepsilon\right) &= P\left(X_n > c + \varepsilon\right) + P\left(X_n < c - \varepsilon\right) \\ &\le 1 - F_n(c + \varepsilon) + F_n(c - \varepsilon) \to 1 - 1 + 0 = 0 \end{aligned}$$

as $n \to \infty$. Note that as $c + \varepsilon > c$ and $c - \varepsilon < c$, $F_n(c + \varepsilon) \to 1$ and $F_n(c - \varepsilon) \to 0$. Thus, $X_n \to c$ in probability. □

Corollary 4.1. *Suppose $\{X_n\}$ is a sequence of RVs defined on a probability space. Then, $X_n \to c$ in distribution if and only if $X_n \to c$ in probability, where c is a constant.*

Proof: The proof of the corollary is straight forward by combining the previous theorem and Theorem 4.3. □

The following theorems provide several properties of different modes of convergence. The proof of the theorems are skipped here.

Theorem 4.5. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Suppose $X_n \to X$ w. p. 1 and $Y_n \to Y$ w. p. 1. Then*

- *$X_n + Y_n \to X + Y$ w. p. 1.*
- *$X_nY_n \to XY$ w. p. 1.*
- *$f(X_n) \to f(X)$ w. p. 1, for any f continuous.*

Theorem 4.6. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Suppose $X_n \to X$ in probability and $Y_n \to Y$ in probability. Then*

- *$X_n + Y_n \to X + Y$ in probability.*
- *$X_nY_n \to XY$ in probability.*
- *$f(X_n) \to f(X)$ in probability, for any f continuous.*

Theorem 4.7. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$.*

- *If $X_n \to X$ in r^{th} mean and $Y_n \to Y$ in r^{th} mean, then $X_n + Y_n \to X + Y$ in r^{th} mean.*
- *If $X_n \to X$ in r^{th} mean then $f(X_n) \to f(X)$ in r^{th} mean, for any f bounded continuous.*

Theorem 4.8. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Suppose $X_n \to X$ in distribution and $Y_n \to c$ in probability for some constant c. Then*

- *$X_n + Y_n \to X + c$ in distribution.*
- *$X_nY_n \to cX$ in distribution.*
- *$f(X_n) \to f(X)$ in distribution, for any f continuous.*

Example 4.10. Let X, $Y \sim N(0,\, 1)$ and X and Y be independent RVs. Take $X_n = X$ and $Y_n = Y$ for all $n = 1,\, 2,\, 3,\, \ldots$. Then, $X_n \to X$ in distribution and $Y_n \to X$ in distribution. Now, $X_n + Y_n = X + Y \sim N(0,\, 2)$ and $2X \sim N(0,\, 4)$. Thus, $X_n + Y_n$ does not converges to $2X$ in distribution. This example shows that $X_n + Y_n$ may not converge to $X + Y$ in distribution if $X_n \to X$ in distribution and $Y_n \to Y$ in distribution. You can easily check that the same conclusion is also true for product. ||

Theorem 4.9. *Let X_n be a RV with MGF $M_n(t)$ for $n = 1,\, 2,\, 3,\, \ldots$. Let X be a RV with MGF $M(t)$. If $M_n(t) \to M(t)$ for all t in an open interval containing zero, as $n \to \infty$, then $X_n \to X$ in distribution.*

Theorem 4.10. *Let X_n be a DRV with PMF $f_n(\cdot)$ for $n = 1, 2, 3, \ldots$. Let X be a DRV with PMF $f(\cdot)$. If, for all $x \in \mathbb{R}$, $f_n(x) \to f(x)$ as $n \to \infty$, then $X_n \to X$ in distribution.*

Theorem 4.11. *Let X_n be a CRV with PDF $f_n(\cdot)$ for $n = 1, 2, 3, \ldots$. Let X be a CRV with PDF $f(\cdot)$. If, for all $x \in \mathbb{R}$, $f_n(x) \to f(x)$ as $n \to \infty$, then $X_n \to X$ in distribution.*

Example 4.11. Let $X_n \sim Bin(n, p_n)$, where $p_n \to 0$ and $np_n = \lambda\,(> 0)$. Then, for $n = 1, 2, 3, \ldots$, the MGF of X_n is

$$M_n(t) = \left(1 - p_n + p_n e^t\right)^n = \left(1 + \frac{\lambda}{n}\left(e^t - 1\right)\right)^n \to e^{\lambda\left(e^t-1\right)}$$

for all $t \in \mathbb{R}$. Note that if $X \sim Poi(\lambda)$, then the MGF of X is

$$M(t) = e^{\lambda\left(e^t-1\right)} \quad \text{for } t \in \mathbb{R}.$$

Thus, $X_n \to X$ in distribution.

Recall that the motivation of the Poission distribution was not discussed when it was introduced. This example tells us the motivation behind the Poission distribution. We can use Poisson distribution to approximate the probability of a Binomial distribution when probability of success is very small and number of trials is very large. ||

Example 4.12. Under the conditions of the previous example, we can prove that $X_n \to X$ using Theorem 4.10. To see it, we can proceed as follows.

$$\begin{aligned}
P\left(X_n = k\right) &= \binom{n}{k} p_n^k \left(1 - p_n\right)^{n-k} \\
&= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \times \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\
&\to \frac{e^{-\lambda}\lambda^k}{k!}.
\end{aligned}$$

Notice that the support of X_n is the set $\{0, 1, 2, \ldots, n\}$. When $n \to \infty$, the support becomes $\{0, 1, 2, \ldots\}$. ||

Example 4.13. Let $X_n \sim U(0, 1 + 1/n)$ for $n = 1, 2, 3, \ldots$. Then the PDF of X_n is

$$f_n(x) = \begin{cases} \frac{1}{1+\frac{1}{n}} & \text{if } 0 < x < 1 + \frac{1}{n} \\ 0 & \text{otherwise} \end{cases} \longrightarrow f(x) = \begin{cases} 1 & \text{if } 0 < x \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

which is the PDF of a RV X such that $X \sim U(0, 1)$. Thus, $X_n \to X$ in distribution. ||

4.2 Limit Theorems

In this section, we will discuss two very famous and useful theorems. Again, we will skip the proofs, but we will see some applications.

Theorem 4.12 (Strong Law of Large Numbers)**.** *Let $\{X_n\}$ be a sequence of i.i.d. RVs with finite mean μ. Define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\{\overline{X}_n\}$ converges to μ almost surely.*

Proof: The proof is skipped. □

Let us loosely discuss the intuitive idea of the previous theorem. Suppose that we want to find the average height of all Indians. Ideally, we need to go to each and every Indian and record their height. Finally, the average should be calculated based on the observations on height. This average is called population average or population mean. It is a very costly (in terms of money and time) process. Alternatively, we can take a representative sample of the Indian population. Here, sample represents a subset of original population. Then, we can collect the height data for each and every person in the sample and then calculate the mean of those sample observations. This mean is called sample mean. If the number of persons in the sample is very small (say, 5 or 10), the calculated sample mean may not be close to the original population mean. However, if we keep on increasing the sample size (the number of persons in the sample), the sample mean should get closer to population mean. The above theorem provided theoretical justification of this intuitive idea. Note that μ and $\overline{X}$ are population and sample means, respectively. Thus, loosely speaking, the strong law of large numbers (SLLN) states that sample mean converges to population mean almost surely as we increase the sample size.

Example 4.14 (Bernoulli proportion converges to success probability)**.** Suppose that a sequence of independent trials is performed. Let E be a fixed event. Letting

$$X_i = \begin{cases} 1 & \text{if } E \text{ occurs on the } i\text{th trial} \\ 0 & \text{if } E \text{ does not occur on the } i\text{th trial,} \end{cases}$$

we have by the SLLN that, with probability one,

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu = E(X_1) = P(E).$$

Since, $X_1 + X_2 + \ldots + X_n$ represents the number of times that the event E occurs in the first n trials, we may interpret it as stating that, with probability one, the limiting proportion of time that the event E occurs is $P(E)$. ||

Example 4.15 (Monte Carlo Integration)**.** Suppose that we want to integrate

$$I = \int_a^b h(x)dx.$$

If we cannot do it explicitly, we can use numerical technique like Simpson's 1/3rd rule. Here, we will see another technique based on the SLLN. Suppose that a and b are finite real numbers. Note that the above integration can be rewritten as

$$I = (b-a)\int_a^b h(x)\frac{1}{b-a}dx = (b-a)E\left(Y\right),$$

where $Y = h(X)$ and $X \sim U(a,\, b)$. Let $\{X_n\}$ be a sequence of *i.i.d.* RVs with common distribution $U(a,\, b)$ and assume that $Y_n = h(X_n)$ for $n = 1,\, 2,\, 3,\, \ldots$. Now, SLLN says that, with probability one,

$$\overline{Y}_n = \frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \to E(Y) = \frac{I}{b-a} \implies \frac{b-a}{n}\sum_{i=1}^{n} h(X_i) \to I.$$

Thus, we can generate N random numbers from $U(a,\, b)$. The generation from $U(a,\, b)$ can be done using any standard software like R, MATLAB, etc. Here, N is a large integer (the popular choices are 5000 or 10000). Then, the integration I can be approximated using $\frac{b-a}{N}\sum_{i=1}^{N} h(X_i)$. ||

Theorem 4.13 (Central Limit Theorem). *Let $\{X_n\}$ be a sequence of i.i.d. RVs with mean μ and variance $\sigma^2 < \infty$. Then, as $n \to \infty$,*

$$P\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \le a\right) \to \Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Proof: The proof is skipped. □

The central limit theorem (CLT) says that

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \to Z \sim N(0,\, 1) \quad \text{in distribution.}$$

Thus, the CDF of standardized sample mean can be approximated (for large sample size) using the CDF of a standard normal distribution, whenever X_n's are i.i.d. RVs with finite mean μ and finite variance σ^2. In other words, the CDF of sample mean can be approximated using the CDF of a $N(\mu,\, \frac{\sigma^2}{n})$ distribution. Note that CLT holds true for any distribution of X_n as long as the variance is finite.

Example 4.16 (Normal Approximation to the Binomial). Let $X_n \sim Bin(n,\, p)$. Then

$$P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \le a\right) \to \Phi(a) \quad \text{as } n \to \infty.$$

We will use CLT to prove this statement. Let $\{Y_n\}$ be a sequence of i.i.d. RVs where $Y_1 \sim Bernoulli(p)$. Then, we know that

$$\sum_{i=1}^{n} Y_i \overset{d}{=} X_n \implies \overline{Y}_n \overset{d}{=} \frac{X_n}{n}.$$

Now, $E(Y_n) = p$ and $Var(Y_n) = p(1-p)$ for all $n = 1,\, 2,\, 3,\, \ldots$. Thus,

$$P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \le a\right) = P\left(\sqrt{n}\frac{\overline{Y}_n - p}{\sqrt{p(1-p)}} \le a\right) \to \Phi(a) \quad \text{as } n \to \infty.$$

The equality in the above line is due to the fact that $\overline{Y}_n$ and $\frac{X_n}{n}$ have same distribution. The convergence is due to the CLT. ||

Example 4.17. The lifetimes of a special type of battery is a RV with mean 40 hours and standard deviation 20 hours. A battery is used until it fails, at which point it is replaced by a new one. Assume a stockpile of 25 such batteries, the lifetimes of which are independent, we want to approximate the probability that over 1100 hours of use can be obtained. Let X_i denote the lifetime of the ith battery to be put in use. Then, we are interested in

$$p = P\left(X_1 + X_2 + \ldots + X_{25} > 1100\right),$$

which can be approximated as follows:

$$\begin{aligned}
p &= P\left(X_1 + X_2 + \ldots + X_{25} > 1100\right) \\
&= P\left(\overline{X}_{25} > 44\right) \\
&= P\left(\sqrt{25}\,\frac{\overline{X}_{25} - 40}{20} > \sqrt{25}\,\frac{44 - 40}{20}\right) \\
&\approx P\left(Z > 1\right), \text{ where } Z \sim N(0,\,1). \text{ This is due to CLT} \\
&= 1 - \Phi(1) \approx 0.1587,
\end{aligned}$$

as $\Phi(1) \approx 0.8413$. This values can be found from the normal table. ||