1. Probability Notes

In this course, we are going to study core concepts of Probability, which shall help us in analyzing 'randomness'/uncertainty in real world situations.

Example 1.1 (Toss a coin)**.** If we toss a coin, then either a head or a tail will appear. For simplicity, we do not consider the unlikely event in which the coin lands on its edge. Here the results/outcomes are non-numerical.

Example 1.2 (Throw a die)**.** If we throw/roll a standard six-sided die and observe the number that appears on the top, then we would have one of the numbers $1, 2, 3, 4, 5, 6$ as the result/outcome. Here, the outcome is numerical and the values are in $\{1, 2, 3, 4, 5, 6\}$.

Example 1.3 (Lifetime of an electric bulb)**.** Switch on a new electric bulb and wait till the time it fails. The duration in which the bulb was working gives us the lifetime of the bulb. The result/outcome is some real number in $[0, \infty)$.

Remark 1.4 (What is Probability?). Probability theory is a branch of Pure Mathematics, and deals with objects involving 'randomness'/uncertainty. As in Pure Mathematics, certain Axioms/hypotheses shall be assumed and results shall be derived from these assumptions. However, it turns out that in many real world situations we may take appropriate models in probability and it will replicate the intrinsic features from the real world – in this sense, Probability may also be considered as 'applicable'. In Example 1.3, probability may represent the law according to which the lifetimes vary across multiple electric bulbs.

Remark 1.5 (What is Statistics?). In Statistics we are faced with data/sample from an underlying population (for example, consider the lifetimes of 5 electric bulbs from a batch of 100 bulbs in Example 1.3). These data/sample typically consists of measurements in an experiment, responses in a survey etc.. We would like to make various kinds of inferences (involving characteristics) about the underlying population from the data/sample provided. We are also interested in the procedures through which such an analysis may be done and the effectiveness of such procedures. These topics are not part of this course.

Note 1.6. A reasonable approach in studying any new random phenomena is to perform experiments under controlled situations, by repeating the phenomena under identical conditions. After a sufficient number of repetitions, we may have some idea about outcomes/'events' which are more likely to occur than other such outcomes/events. We must note, however, that each experiment terminates in an outcome, which cannot be specified in advance, i.e. before performing the experiment.

Definition 1.7 (Random experiment)**.** A random experiment is an experiment in which

(a) all possible outcomes of the experiment are known in advance,
(b) outcome of a particular trial/performance of the experiment cannot be specified in advance,
(c) the experiment can be repeated under identical conditions.

Notation 1.8. A random experiment shall be denoted by $\mathcal{E}$.

Definition 1.9 (Sample Space)**.** The collection of all possible outcomes of a random experiment $\mathcal{E}$ is called its sample space.

Notation 1.10. A sample space shall be denoted by Ω. It is a set containing all possible outcomes.

Example 1.11 (Examples of Random experiments and corresponding Sample spaces)**.** The experiments mentioned in Examples 1.1, 1.2 and 1.3 are all examples of random experiments. The corresponding sample spaces are $\{H, T\}, \{1, 2, 3, 4, 5, 6\}$ and $[0, \infty)$ respectively. Here, H and T denotes a head and a tail respectively.

Example 1.12 (Tossing two coins simultaneously)**.** If we write the result/outcome from the first coin as x and the second coin as y, then the result of the experiment may be written as an ordered pair (x, y). Here, x is either a head or a tail. Similarly, y is either a head or a tail. The sample space is therefore,

$$\Omega = \{(x, y) : x, y \in \{H, T\}\} = \{(H, H), (H, T), (T, H), (T, T)\}.$$

Notation 1.13. Given two sets A and B, we write $A \times B := \{(x, y) : x \in A, y \in B\}$. In Example 1.12, $\Omega = \{H, T\} \times \{H, T\}$. The set $\mathbb{R}^2$ is nothing but $\mathbb{R} \times \mathbb{R}$.

Example 1.14 (Throwing a die three times)**.** In this case, we record the outcome of all three throws taken together. If x, y and z represent the result/outcome of the first, second and the third throws respectively, then the outcome may be represented as the ordered triple (x, y, z). The sample space is therefore

$$\Omega = \{(x, y, z) : x, y, z \in \{1, 2, 3, 4, 5, 6\}\}.$$

Note 1.15. We are interested in specific outcomes or more generally, specific subsets of the sample space Ω, which are more likely to appear than other such subsets. In the case where we deal with specific outcomes, we shall consider them as singleton subsets of Ω.

Definition 1.16 (Events)**.** If the outcome of a random experiment $\mathcal{E}$ is an element of a subset E of Ω, then we say that the event E has occured.

Notation 1.17. As mentioned in the previous definition, we are interested in specific subsets of Ω, to be referred to as events. The collection of all events shall be denoted as $\mathcal{F}$.

Note 1.18. The empty set $\emptyset$ and the sample space Ω will always be an element in $\mathcal{F}$.

Remark 1.19. In many situations, we shall take the event space $\mathcal{F}$ as the power set 2^Ω of Ω. Recall that the power set of Ω is the collection of all subsets of Ω. Later on, we shall discuss specific situations in which we may restrict our attention to a smaller collection than 2^Ω.

Notation 1.20. We may refer to a collection of sets as a class of sets. The event space $\mathcal{F}$ is a class of subsets of the sample space Ω.

Example 1.21 (Examples of Events)**.**

(a) $\{H\}$ and $\{T\}$ are events in Example 1.1. The event space $\mathcal{F}$ may be taken as $\mathcal{F} = 2^\Omega = \{\emptyset, \{H\}, \{T\}, \Omega\}$ with $\Omega = \{H, T\}$.

(b) $\{4\}, [5, \infty), [2, 3], [1, 100)$ are events in Example 1.3.

(c) $\{(1, 4, 5), (2, 2, 2), (3, 6, 2)\}$ is an event in Example 1.14.

Remark 1.22. Observe that complementation of an event E gives us the subset E^c. The set E^c may be interpretated as the non-occurrence of the event E. Thus, we treat E^c as another event.

Similarly, finite or countably infinite unions and intersections of events give us further events. Therefore, we can consider standard set theoretic operations, viz. complementation, finite and countably infinite unions and intersections on the event space $\mathcal{F}$.

Note 1.23. For technical reasons, we do not consider uncountable unions or intersections of events.

Remark 1.24. As mentioned earlier in Note 1.15, we would like to identify special subsets or events which are more likely to occur than the others. This is where Probability enters the discussion. Probability is a measure of uncertainty and we are interested in associating numerical quantities to events/outcomes thereby quantifying the uncertainty related to these events/outcomes. This is achieved by assigning probabilities to the events.

Definition 1.25 (A priori or Classical definition of probability)**.** Suppose that a random experiment results in n (a finite number) outcomes. Given an event $A \in \mathcal{F}$, if it appears in m $(0 \leq m \leq n)$ outcomes, then the probability of A is $\frac{m}{n}$.

Note 1.26. Given a random experiment, we are already aware of all possible outcomes. Therefore, without performing the experiment, we can discuss about the a priori definition of probability.

Note 1.27. The classical definition works only when there are finitely many outcomes. Due to the limitations of this definition, we look for other ways to understand the notion of probability.

Remark 1.28 (A posteriori or Relative Frequency definition of probability). If a random experiment $\mathcal{E}$ is repeated a large number, say n, of times and an event A occurs m many times, then the relative frequency $\frac{m}{n}$ may be taken as an approximate value of the probability of A. This concept of probability assumes that in the long run there is a regularity of occurence of the event.

Note 1.29. The a posteriori definition of probability works only after performing the random experiment.

Note 1.30. We now discuss an axiomatic definition of probability. We shall recover the classical definition as part of the axiomatic definition and also justify the relative frequency definition of probability.

Note 1.31. At this moment, we do not focus on how the probabilities of events are assigned, i.e. how a probability model is developed. Our interest is in the properties of probability as a measure of uncertainty/'randomness'.

Definition 1.32 (Set function)**.** A set function is a function whose domain is a collection/class of sets.

Definition 1.33 (Probability function/measure)**.** Suppose that Ω and $\mathcal{F}$ are the sample space and the event space of a random experiment $\mathcal{E}$ respectively. A real valued set function $\mathbb{P}$, defined on the event space $\mathcal{F}$, is said to be a probability function/measure if it satisfies the following axioms/properties, viz.

(a) $\mathbb{P}(\Omega) = 1$.

(b) (non-negativity) $\mathbb{P}(E) \geq 0$ for any event E in $\mathcal{F}$.

(c) (Countable additivity) If $\{E_n\}_n$ is a sequence of events in $\mathcal{F}$ such that $E_i \cap E_j = \emptyset, \forall i \neq j$, then $\mathbb{P}(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} \mathbb{P}(E_n)$.

Definition 1.34 (Probability space)**.** If $\mathbb{P}$ is a probability function defined on the event space $\mathcal{F}$ of a random experiment $\mathcal{E}$, then the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be a probability space. Here, Ω denotes the sample space of $\mathcal{E}$.

Definition 1.35 (Mutually Exclusive or Pairwise disjoint events)**.** Let $\mathcal{I}$ be an indexing set. A collection of events $\{E_i : i \in \mathcal{I}\}$ is said to be mutually exclusive or pairwise disjoint if $E_i \cap E_j = \emptyset, \forall i \neq j$.

Note 1.36. We first study some basic properties of probability functions and then look at some explicit examples.

Proposition 1.37. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space associated with a random experiment $\mathcal{E}$.*

(a) $\mathbb{P}(\emptyset) = 0$.

Proof. Take the sequence of events $\{E_n\}_n$ in $\mathcal{F}$ given by $E_1 := \Omega$ and $E_n = \emptyset, \forall n \geq 2$. Then $\bigcup_n E_n = \Omega$ and the E_n's are pairwise disjoint. By Definition 1.33,

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\bigcup_n E_n) = \sum_{n=1}^{\infty} \mathbb{P}(E_n) = \mathbb{P}(E_1) + \sum_{n=2}^{\infty} \mathbb{P}(E_n) = 1 + \sum_{n=2}^{\infty} \mathbb{P}(E_n).$$

Therefore,

$$0 = \sum_{n=2}^{\infty} \mathbb{P}(E_n) = \lim_{m\to\infty} \sum_{n=2}^{m} \mathbb{P}(E_n) = \lim_{m\to\infty} [(m-1)\mathbb{P}(\emptyset)].$$

The result follows. □

(b) (Finite additivity) Let $E_1, E_2, \cdots, E_n \in \mathcal{F}$ for some integer $n \geq 2$ be mutually exclusive events. Then $\mathbb{P}(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n \mathbb{P}(E_i)$.

Proof. Consider the events $E_i = \emptyset, \forall i > n$. Then $\mathbb{P}(E_i) = 0, \forall i > n$ and the sequence of events $\{E_m\}_m$ is mutually exclusive. Now,

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

□

(c) $\mathbb{P}(E) + \mathbb{P}(E^c) = 1$ for all events $E \in \mathcal{F}$.

Proof. Note that $E \cap E^c = \emptyset$, i.e. the events E and E^c are mutually exclusive. Then by finite additivity, $\mathbb{P}(E) + \mathbb{P}(E^c) = \mathbb{P}(E \bigcup E^c) = \mathbb{P}(\Omega) = 1$. □

(d) $0 \leq \mathbb{P}(E) \leq 1$ for all events E in $\mathcal{F}$.

Proof. The inequality $\mathbb{P}(E) \geq 0$ follows from the definition. Again $\mathbb{P}(E^c) \geq 0$. Using $\mathbb{P}(E) + \mathbb{P}(E^c) = 1$, we have $\mathbb{P}(E) \leq \mathbb{P}(E) + \mathbb{P}(E^c) = 1$. □

(e) (Monotonicity) Suppose $A, B \in \mathcal{F}$ with $A \subseteq B$. Then $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$. In particular, $\mathbb{P}(A) \leq \mathbb{P}(B)$. If, in addition $\mathbb{P}(B) = 0$, then $\mathbb{P}(A) = 0$.

Proof. Observe that the sets A and $A^c \cap B$ are mutually exclusive and that $B = A \bigcup (A^c \cap B)$. By finite additivity, we have $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$. As $\mathbb{P}(A^c \cap B) \geq 0$, hence $\mathbb{P}(A) \leq \mathbb{P}(B)$.

If $\mathbb{P}(B) = 0$, then $0 \leq \mathbb{P}(A) \leq \mathbb{P}(B) = 0$. Hence, $\mathbb{P}(A) = 0$. □

(f) *(Inclusion-Exclusion principle for two events) For $A, B \in \mathcal{F}$, we have*

$$\mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Proof. Observe that $A = (A \cap B) \bigcup (A \cap B^c)$ and the events $A \cap B, A \cap B^c$ are mutually exclusive. Then

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c).$$

Similarly, $B = (A \cap B) \bigcup (A^c \cap B)$ and hence

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B).$$

Then,

$$\mathbb{P}(A) + \mathbb{P}(B) = \mathbb{P}(A \cap B) + [\mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B)].$$

Observe that the events $A \cap B^c, A \cap B, A^c \cap B$ are mutually exclusive and

$$(A \cap B^c) \bigcup (A \cap B) \bigcup (A^c \cap B) = A \bigcup B.$$

Then $\mathbb{P}(A \bigcup B) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B)$ and hence

$$\mathbb{P}(A) + \mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A \bigcup B).$$

The result follows. □

(g) *(Boole's inequality for two events) For $A, B \in \mathcal{F}$, we have $\mathbb{P}(A \bigcup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.*

Proof. Since $\mathbb{P}(A \cap B) \geq 0$, using the Inclusion-Exclusion principle, we have $\mathbb{P}(A \bigcup B) \leq \mathbb{P}(A \cap B) + \mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B)$. □

(h) *(Bonferroni's inequality for two events) For $A, B \in \mathcal{F}$, we have $\mathbb{P}(A \cap B) \geq \max\{0, \mathbb{P}(A) + \mathbb{P}(B) - 1\}$.*

Proof. By definition, we have $\mathbb{P}(A \cap B) \geq 0$. Again, using $\mathbb{P}(A \bigcup B) \leq 1$ and the Inclusion-Exclusion principle, we have $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \bigcup B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$. The result follows. □

Example 1.38 (Probability space associated with a coin toss)**.** Recall from Example 1.21 that in the random experiment of tossing a coin, we have the sample space $\Omega = \{H, T\}$ and the event space $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$. If $\mathbb{P}$ is a probability function defined on $\mathcal{F}$, then we have $\mathbb{P}(\Omega) = 1, \mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\{H\}) + \mathbb{P}(\{T\}) = 1$. The last relation follows from the observation that $\{T\} = \{H\}^c$. If $\mathbb{P}(\{H\}) = p \in [0, 1]$, then $\mathbb{P}(\{T\}) = 1 - p$. These are necessary conditions derived from the axioms/properties. Now we can ask the following: given a function $\mathbb{P}$ on $\mathcal{F}$ defined by

$$\mathbb{P}(\emptyset) := 0,\ \mathbb{P}(\{H\}) := p,\ \mathbb{P}(\{T\}) := 1 - p,\ \mathbb{P}(\Omega) := 1$$

is $\mathbb{P}$ a probability function for any $p \in [0, 1]$? If you have a fair coin, you would expect that the probability of occurrence of a head and a tail should be the same – in which case we have $p = 1 - p$, i.e. $p = \frac{1}{2}$.

Note 1.39. In the next week, we shall see further examples.

Note 1.40. In the examples discussed in the previous week, we have the corresponding sample spaces are either finite or uncountably infinite.

Example 1.41 (Throw/Roll a die until 6 appears)**.** Suppose we take a standard six-sided die and count the number of rolls required to obtain the first 6. In this case, our sample space is $\Omega = \{1, 2, \cdots\}$, which is countably infinite.

Remark 1.42. If $(\Omega, \mathcal{F} = 2^\Omega, \mathbb{P})$ is a probability space, with Ω being a finite or a countably infinite set, then all its subsets $A \in \mathcal{F}$ are also finite or countably infinite. By finite/countable additivity of $\mathbb{P}$, we have

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}), \forall A \in \mathcal{F}.$$

Note 1.43. In the discussion below, we consider a set Ω, assumed to be a finite or a countably infinite set and discuss structural properties of probability spaces on Ω. The observations here are then applicable to situations where we have a random experiment with only finitely many or countably infinite many outcomes, i.e. the sample space is finite or countably infinite. We are going to see that specifying the probability of singleton events can describe the probability function/measure on the event space $\mathcal{F}$.

Let Ω be any finite or countably infinite set. Consider $\mathcal{F} = 2^{\Omega}$ the power set. Let $p : \Omega \to [0,1]$ be a function such that

$$\sum_{\omega \in \Omega} p_\omega = 1.$$

Now consider the real valued set function $\mathbb{P}$ on $\mathcal{F}$ defined by

$$\mathbb{P}(A) = \sum_{\omega \in A} p_\omega, \forall A \in \mathcal{F}.$$

Note 1.44. Observe that $\mathbb{P}(\{\omega\}) = p_\omega, \forall \omega \in \Omega$.

Proposition 1.45. *The set function $\mathbb{P}$, as defined above, is a probability function/measure on $\mathcal{F}$.*

Proof. We verify the axioms in Definition 1.33. By definition, $\mathbb{P}(\Omega) = \sum_{\omega \in \Omega} p_\omega = 1$.

Since, $p_\omega \geq 0, \forall \omega \in \Omega$, we have $\mathbb{P}(A) = \sum_{\omega \in A} p_\omega \geq 0, \forall A \in \mathcal{F}$.

Let $\{A_n\}_n$ be a sequence of mutually exclusive events. Then each element of $\bigcup_n A_n$ belongs to exactly one A_n. Then,

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} \sum_{\omega \in A_n} p_\omega = \sum_{\omega \in \bigcup_{n=1}^{\infty} A_n} p_\omega = \mathbb{P}(\bigcup_{n=1}^{\infty} A_n).$$

This completes the proof. □

Definition 1.46 (Discrete Probability spaces)**.** Let Ω be a finite or countable set. We refer to a probability space of the form $(\Omega, 2^{\Omega}, \mathbb{P})$ as a discrete probability space.

Remark 1.47. A Probability function/measure on a discrete probability space is determined by the probability of singleton events.

Notation 1.48. We may refer to the singleton events in a discrete probability space as elementary events.

Example 1.49 (Examples of discrete probability spaces)**.** The following are some examples of discrete probability spaces. Here we only specify the probability of singleton sets/events.

(a) $\Omega = \{H, T\}$ with $\mathbb{P}(\{H\}) = p, \mathbb{P}(\{T\}) = 1 - p$ for some fixed $p \in [0,1]$.

(b) $\Omega = \{1, 2, 3, 4, 5, 6\}$ with $\mathbb{P}(\{i\}) = \frac{1}{6}, \forall i \in \Omega$.

(c) Let Ω be the set of all natural numbers, i.e. $\Omega = \{1, 2, 3, \cdots\}$. Take $\mathbb{P}(\{n\}) = \frac{1}{2^n}, \forall n \in \Omega$.

Remark 1.50 (Equally likely probability models on finite sample spaces). Let $\mathcal{E}$ be a random experiment with the sample space $\Omega = \{\omega_1, \cdots, \omega_k\}$, a finite set with k elements. Here, any probability function/measure $\mathbb{P}$ on $\mathcal{F} = 2^\Omega$ is determined by the values $p_{w_i} = \mathbb{P}(\{\omega_i\}), i = 1, \cdots, k$. Assume that the elementary events $\{\omega_i\}$ are equally likely, i.e. $p_{w_i} = \mathbb{P}(\{\omega_i\}) = p_{w_j} = \mathbb{P}(\{\omega_j\}), \forall i \neq j$. Since, $\sum_{\omega \in \Omega} p_\omega = 1$, we have $p_{w_i} = \mathbb{P}(\{\omega_i\}) = \frac{1}{k}, \forall i = 1, \cdots, k$. For any set/event $A \in \mathcal{F}$, we have

$$\mathbb{P}(A) = \sum_{\omega \in A} p_\omega = \frac{\#A}{k},$$

where $\#A$ denotes the cardinality of A, i.e. the number of elements in A. We can rewrite the above observation in terms of the following interpretation.

$$\mathbb{P}(A) = \frac{\text{number of ways favourable to the event } A}{\text{number of ways in which the random experiment can terminate}}.$$

Definition 1.51 (At random)**.** Let $\mathcal{E}$ be a random experiment with finite sample space. We say that the experiment has been performed at random to imply that all the elementary/singleton events are equally likely. Identifying singleton events with the corresponding outcomes, we may also say that the outcomes are equally likely. In this case, the number of ways in which the random experiment can terminate is exactly the cardinality of the sample space.

Note 1.52. While tossing a coin or rolling a die, if the outcomes are equally likely, then we say that the coin/die is 'fair'.

Example 1.53. Example 1.49(b) has been performed at random.

Example 1.54. A box contains 3 red balls and 2 green balls. Balls of the same colour are assumed to be identical. Draw a ball at random. If A denotes the event that the ball drawn is red, then $\mathbb{P}(A) = \frac{3}{5}$.

Note 1.55. Consider a random experiment $\mathcal{E}$ with the sample space Ω being the set of all natural numbers, i.e. $\Omega = \{1, 2, 3, \cdots\}$. Then for any probability function/measure $\mathbb{P}$ on $\mathcal{F} = 2^\Omega$, we have $1 = \mathbb{P}(\Omega) = \sum_{n=1}^\infty p_n, \forall A \in \mathcal{F}$. Consequently, $\lim_n p_n = 0$ and all p_n's cannot be equal. Hence we cannot have natural numbers drawn at random. By a similar argument, we cannot have any random experiment performed at random if the sample space is countably infinite.

Remark 1.56. When multiple draws from a box are involved in a single trial of a random experiment, then there are two broad categories of problems, viz. sampling with replacement and sampling without replacement. In the first case, the outcome of each draw is returned to the box before the next draw. In the second case, the outcome is removed from the possibilities in the next draw. Following examples illustrate these concepts.

Example 1.57. Example 1.14 where we throw/roll a die thrice is an example of sampling with replacement. The cardinality of the sample space is $6 \times 6 \times 6 = 6^3$. If A denote the event that all the rolls result in an even number, then number of ways favourable to A is $3 \times 3 \times 3 = 3^3$. Thus, $\mathbb{P}(A) = \frac{3^3}{6^3} = \frac{1}{8}$.

Example 1.58. Draw 2 cards at random from a standard deck of 52 cards. Here, the cardinality of the sample space is $\binom{52}{2}$. Since, we are looking at the 2 cards in hand together, the order in which they have been obtained does not matter. Consider the event A that both cards are from the Club ($\clubsuit$) suit. Since a standard deck of cards contain 13 cards from the Club suit, we have $\mathbb{P}(A) = \binom{13}{2}/\binom{52}{2}$. This is an example of sampling without replacement.

Example 1.59 (Placing r balls in m bins)**.** Fix two positive integers r and m. Suppose that there are r labelled balls and m labelled bins/boxes/urns. Assume that each bin can hold all the balls, if required. One by one, we put the balls into the bins 'at random'. Then, by letting ω_i be the bin-number into which the i-th ball is placed, we can capture the full configuration by the vector $\underline{\omega} = (\omega_1, \ldots, \omega_r)$. Let Ω be the list of all configurations. Therefore, Ω is the sample space of this random experiment. We have

$$\Omega = \{\underline{\omega} : \underline{\omega} = (\omega_1, \ldots, \omega_r) \text{ with } 1 \leq \omega_i \leq m \text{ for each } 1 \leq i \leq r\}.$$

The cardinality of Ω is m^r (since each ball may be placed in one of the m bins). Since the experiment has been performed at random, we have $\mathbb{P}(\{\underline{\omega}\}) = p_{\underline{\omega}} = m^{-r}, \forall \underline{\omega} \in \Omega$. We now consider the probabilities of the following events.

(a) Let A be the event that the r-th ball is placed in the first bin. Then $A = \{\underline{\omega} \in \Omega : \omega_r = 1\}$. Here, balls numbered 1 to $r-1$ can be placed in any of the m bins. Therefore, the number of outcomes $\underline{\omega}$ favourable to A is m^{r-1}. Hence, $\mathbb{P}(A) = \frac{m^{r-1}}{m^r} = \frac{1}{m}$.

(b) Let B be the event that the first bin is empty. Then $B = \{\underline{\omega} \in \Omega : \omega_i \neq 1, \forall i = 1, 2, \cdots, r\}$. Here, each ball can be placed in any of the remaining bins numbered 2 to m. Since there are $m-1$ choices for each ball, the number of outcomes $\underline{\omega}$ favourable to B is $(m-1)^r$. Hence $\mathbb{P}(B) = \frac{(m-1)^r}{m^r}$.

(c) Consider $r \leq m$ and let C be the event that all the balls are placed in distinct bins, i.e. no bins contain more than one ball (a bin may remain empty). Then, $C = \{\underline{\omega} \in \Omega : \omega_i \neq \omega_j, \forall i \neq j\}$. Here, we are choosing/sampling bins for each ball and the sampling is being done without replacement. Hence, the number of outcomes $\underline{\omega}$ favourable to C is ${}^mP_r$. Hence $\mathbb{P}(C) = {}^mP_r \, m^{-r} = \frac{m(m-1)\cdots(m-r+1)}{m^r} = \frac{(m-1)\cdots(m-r+1)}{m^{r-1}}$.

Example 1.60 (Birthday Paradox)**.** There are n people at a party. What is the chance that two of them have the same birthday? Assume that none of them was born on a leap year and that days are equally likely to be a birthday of a person. The problem structure remains the same as in the previous balls in bin problem, where the bins are labelled as $1, 2, \ldots, 365$ (days of the year), and the balls are labelled as $1, 2, \ldots, n$ (people). In the notations of the previous example, $r = n$ and $m = 365$. Here, we wish to find the probability of the following event

$$D = \{\underline{\omega} \in \Omega : \omega_i = \omega_j, \text{for some } i \neq j\}.$$

Note that $D = C^c$, where C is as in the previous example. Therefore, $\mathbb{P}(D) = 1 - \mathbb{P}(C) = 1 - \frac{(365-1)\cdots(365-n+1)}{365^{n-1}}$. The reason this is called a 'paradox' is that even for n much smaller than 365, the probability becomes significantly large. For example, $n = 25$ gives $\mathbb{P}(D) > 0.5$.

We now discuss a generalization of the Inclusion-Exclusion principle for two events discussed in Proposition 1.37.

Proposition 1.61 (Inclusion Exclusion Principle)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, \ldots, A_n$ be events. Then,*

$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) = S_{1,n} - S_{2,n} + S_{3,n} - \cdots + (-1)^{n-1} S_{n,n},$$

where

$$S_{k,n} := \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}).$$

Proof. For the case $n = 2$, the result has already been discussed in Proposition 1.37. We prove the result for general n by an application of the principle of Mathematical Induction.

Suppose the result is true for $n = 2, 3, \cdots, k$. We want to establish the result for $n = k + 1$. Using the result for $n = 2$, we have

$$\mathbb{P}(\bigcup_{i=1}^{k+1} A_i) = \mathbb{P}((\bigcup_{i=1}^{k} A_i) \cup A_{k+1}) = \mathbb{P}(\bigcup_{i=1}^{k} A_i) + \mathbb{P}(A_{k+1}) - \mathbb{P}((\bigcup_{i=1}^{k} A_i) \cap A_{k+1}).$$

Consider

$$T_{j,k} := \sum_{1 \leq i_1 < i_2 < \ldots < i_j \leq k} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j} \cap A_{k+1}), \; j = 1, 2, \cdots, k.$$

Applying the result for $n = k$ on the set $\bigcup_{i=1}^{k}(A_i \cap A_{k+1})$, we have

$$\mathbb{P}((\bigcup_{i=1}^{k} A_i) \cap A_{k+1}) = \sum_{j=1}^{k} (-1)^{j-1} T_{j,k}.$$

Then,

$$\mathbb{P}(\bigcup_{i=1}^{k+1} A_i)$$
$$= (S_{1,k} + \mathbb{P}(A_{k+1})) - (S_{2,k} + T_{1,k}) + (S_{3,k} + T_{2,k}) \cdots + (-1)^{k-1}(S_{k,k} + T_{k-1,k}) + (-1)^{(k+1)-1} T_{k,k}$$
$$= \sum_{j=1}^{k+1} (-1)^{j-1} S_{j,k+1}.$$

Hence the result is true for the case $n = k+1$. Applying the principle of Mathematical Induction, the result is true for any positive integer n. □

Example 1.62. Consider placing r labelled balls in m labelled bins at random (Example 1.59). Let E denote the event that at least one bin is empty. Now, for $j = 1, 2, \cdots, m$, consider the event E_j that none of the balls are placed in the j-th bin, i.e. j-th bin is empty. Then

$$E_j = \{\underline{\omega} \in \Omega : \omega_i \neq j, \forall i = 1, 2, \cdots, r\}, \forall j = 1, 2, \cdots, m$$

and $E = \bigcup_{j=1}^m E_j$. Not all the bins can be empty and hence $\bigcap_{j=1}^m E_j = \emptyset$. For $1 \leq k \leq m-1$, $E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_k}$ denotes the event that the bins numbered $j_1 < j_2 < \cdots < j_k$ are empty. Here, each ball can be placed in the remaining $m - k$ bins. Therefore,

$$\mathbb{P}(E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_k}) = \frac{(m-k)^r}{m^r}.$$

By the Inclusion-Exclusion Principle,

$$\mathbb{P}(E) = \mathbb{P}\left(\bigcup_{j=1}^m E_j\right) = \sum_{k=1}^{m-1} (-1)^{k-1} \binom{m}{k} \frac{(m-k)^r}{m^r}.$$

Remark 1.63 (Bonferroni's inequality). In the notations of Proposition 1.61, it can be shown that $S_{1,n} - S_{2,n} \leq \mathbb{P}(\bigcup_{i=1}^n A_i) \leq S_{1,n}$. More generally,

$$\mathbb{P}(\bigcup_{i=1}^n A_i) \leq S_{1,n} - S_{2,n} + \cdots + S_{m,n} \quad \text{if } m \text{ is odd,}$$
$$\mathbb{P}(\bigcup_{i=1}^n A_i) \geq S_{1,n} - S_{2,n} + \cdots - S_{m,n} \quad \text{if } m \text{ is even.}$$

We do not discuss the proof. These inequalities are sometimes referred to as Bonferroni's inequalities in the literature.

Note 1.64. During the performance of a random experiment, if an event A is observed, then it may also provide some information regarding other events. The next concept attempts to formalize this information.

Definition 1.65 (Conditional Probability)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let A be an event with $\mathbb{P}(A) > 0$. For any event B, we define

$$\mathbb{P}(B \mid A) := \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$$

to be the conditional probability of B given A.

Example 1.66. Consider rolling a fair die twice. The sample space is $\Omega = \{(i,j) : i,j \in \{1,2,3,4,5,6\}\}$ with $\mathbb{P}((i,j)) = \frac{1}{36}, \forall (i,j) \in \Omega$. Consider the events $A = \{(i,j) : i \text{ is odd}\}$ and $B = \{(i,j) : i+j = 3\} = \{(1,2),(2,1)\}$. Here, number of outcomes favourable to A is $3 \times 6 = 18$ and hence $\mathbb{P}(A) = \frac{18}{36} = \frac{1}{2} > 0$. The event $A \cap B = \{(1,2)\}$ and hence $\mathbb{P}(A \cap B) = \frac{1}{36}$. Then $\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{1}{18}$.

Note 1.67. We note some basic properties of conditional probability. If $\mathbb{P}(A) > 0$, then

(a) $\mathbb{P}(\Omega \mid A) = \frac{\mathbb{P}(\Omega \,\cap\, A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A)} = 1$.
(b) $\mathbb{P}(A \mid A) = 1$
(c) $B \mapsto \mathbb{P}(B \mid A)$ defines a non-negative set function on $\mathcal{F}$.

Proposition 1.68. $(\Omega, \mathcal{F}, \mathbb{P}(\cdot \mid A))$ *is a probability space.*

Proof. We verify the axioms in Definition 1.33. We have already observed $\mathbb{P}(\Omega \mid A) = 1$ and non-negativity in the above note.

To establish the countable additivity. If $\{E_n\}_n$ is a sequence of mutually exclusive events, then so are $\{E_n \cap A\}_n$. By the countable additivity of $\mathbb{P}$, we have

$$\mathbb{P}(\bigcup_{n=1}^{\infty} E_n \mid A) = \frac{\mathbb{P}(\cup_{n=1}^{\infty} E_n \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(\cup_{n=1}^{\infty} (E_n \cap A))}{\mathbb{P}(A)} = \sum_{n=1}^{\infty} \frac{\mathbb{P}(E_n \cap A)}{\mathbb{P}(A)} = \sum_{n=1}^{\infty} \mathbb{P}(E_n \mid A).$$

This completes the proof. □

Proposition 1.69 (Multiplication Rule)**.** *Let* $E_1, E_2, \cdots, E_n$ *be events in a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *such that* $\mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n) > 0$. *Then*

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

Proof. Note that $E_1 \cap E_2 \cap \cdots \cap E_n \subseteq E_1 \cap E_2 \cap \cdots \cap E_{n-1} \subseteq E_1 \cap E_2 \subseteq E_1$ and hence by the hypothesis, all the conditional probabilities in the statement are well-defined.

For the case $n = 2$, the result follows from the definition of $\mathbb{P}(E_2 \mid E_1)$. For general n, apply result for two events repeatedly in the following manner:

$$
\begin{aligned}
&\mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n) \\
&= \mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_{n-1})\ \mathbb{P}(E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1}) \\
&= \mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_{n-2})\ \mathbb{P}(E_{n-1} \mid E_1 \cap E_2 \cap \cdots \cap E_{n-2})\ \mathbb{P}(E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1}) \\
&= \cdots \\
&= \mathbb{P}(E_1)\ \mathbb{P}(E_2 \mid E_1)\ \mathbb{P}(E_3 \mid E_1 \cap E_2) \cdots\ \mathbb{P}(E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1}).
\end{aligned}
$$

This completes the proof. □

Example 1.70. Suppose that an urn contains 3 red balls and 5 green balls. All balls of the same colour are identical. Suppose 2 balls are drawn successively at random from the urn without replacement. Let A and B denote the events that the first ball is red and second ball is green, respectively. Then $\mathbb{P}(A) = \frac{3}{8}$. If the event A has already happened, then at the time of the second draw the urn contains 2 red balls and 5 green balls. As such $\mathbb{P}(B \mid A) = \frac{5}{7}$. By the Multiplication rule, the probability that the first ball drawn is red and the second ball drawn is green is $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A) = \frac{15}{56}$.

Definition 1.71 (Exhaustive events)**.** Let $\mathcal{I}$ be an indexing set. A collection of events $\{E_i : i \in \mathcal{I}\}$ is said to be exhaustive if $\cup_{i \in \mathcal{I}} E_i = \Omega$.

Theorem 1.72 (Theorem of Total Probability)**.** *Let $\mathcal{I}$ be a finite or countably infinite indexing set. Let $\{E_i : i \in \mathcal{I}\}$ be a collection of mutually exclusive and exhaustive events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{P}(E_i) > 0, \forall i$. Then*

$$\mathbb{P}(E) = \sum_{i \in \mathcal{I}} \mathbb{P}(E \cap E_i) = \sum_{i \in \mathcal{I}} \mathbb{P}(E_i)\ \mathbb{P}(E \mid E_i).$$

Proof. Since E_i's are exhaustive, we have $\mathbb{P}(\cup_{i\in\mathcal{I}}E_i) = \mathbb{P}(\Omega) = 1$. Then $\mathbb{P}(E) = \mathbb{P}(E \cap (\cup_{i\in\mathcal{I}}E_i))$ (see practice problem set 1). Since E_i's are mutually exclusive, so are $E \cap E_i$'s. Hence by the finite or countable additivity of $\mathbb{P}$ (depending on whether $\mathcal{I}$ is finite or countably infinite), we have

$$\mathbb{P}(E) = \mathbb{P}(E \cap (\cup_{i\in\mathcal{I}}E_i)) = \sum_{i\in\mathcal{I}} \mathbb{P}(E \cap E_i) = \sum_{i\in\mathcal{I}} \mathbb{P}(E_i)\ \mathbb{P}(E \mid E_i).$$

This completes the proof. □

Remark 1.73. The practice problem referred in the above proof does not require the fact that $\cup_{i\in\mathcal{I}}E_i = \Omega$, but rather we need $\mathbb{P}(\cup_{i\in\mathcal{I}}E_i) = 1$. In the hypothesis of the previous theorem, we may replace the exhaustiveness of E_i's by the condition $\mathbb{P}(\cup_{i\in\mathcal{I}}E_i) = 1$.

Example 1.74. Suppose we perform a random experiment with the following steps.

(a) Suppose that there are two urns. The first urn contains 3 red balls and 5 green balls. The second urn contains 6 red balls and 3 green balls. All balls of the same colour are identical.

(b) Suppose a fair die is rolled and if the outcome is 1 or 6, then the first urn is chosen. Otherwise, the second urn is chosen.

(c) Finally, 2 balls are drawn at random from the chosen urn.

We want to find the probability that both the balls drawn are red. Let E denote this event. Suppose U_1 and U_2 denote the events that the first urn and the second urn is chosen respectively. Then the events $U_i, i = 1, 2$ are mutually exclusive and exhaustive. Moreover, $\mathbb{P}(U_1) = \frac{2}{6} = \frac{1}{3}$ and $\mathbb{P}(U_2) = \frac{4}{6} = \frac{2}{3}$. Further,

$$\mathbb{P}(E \mid U_1) = \frac{\binom{3}{2}}{\binom{8}{2}} = \frac{3}{28},\ \mathbb{P}(E \mid U_2) = \frac{\binom{6}{2}}{\binom{9}{2}} = \frac{15}{36} = \frac{5}{12}.$$

Then the required probability can be computed as an application of Theorem of Total Probability as

$$\mathbb{P}(E) = \mathbb{P}(U_1)\ \mathbb{P}(E \mid U_1) + \mathbb{P}(U_2)\ \mathbb{P}(E \mid U_2) = \frac{1}{3}\frac{3}{28} + \frac{2}{3}\frac{5}{12} = \frac{1}{28} + \frac{5}{18} = \frac{79}{252}.$$

Theorem 1.75 (Bayes' Theorem)**.** *Let $\mathcal{I}$ be a finite or countably infinite indexing set. Let $\{E_i : i \in \mathcal{I}\}$ be a collection of mutually exclusive and exhaustive events in a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$

such that $\mathbb{P}(E_i) > 0, \forall i$. *Then for any event* $E \in \mathcal{F}$ *with* $\mathbb{P}(E) > 0$, *we have*

$$\mathbb{P}(E_j \mid E) = \frac{\mathbb{P}(E_j)\,\mathbb{P}(E \mid E_j)}{\sum_{i\in\mathcal{I}} \mathbb{P}(E_i)\,\mathbb{P}(E \mid E_i)}, \forall j \in \mathcal{I}.$$

Proof. For any $j \in \mathcal{I}$, we have

$$\mathbb{P}(E_j \mid E) = \frac{\mathbb{P}(E_j \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E_j)\,\mathbb{P}(E \mid E_j)}{\sum_{i\in\mathcal{I}} \mathbb{P}(E_i)\,\mathbb{P}(E \mid E_i)}.$$

In the last step of the calculation above, we have used the Multiplication rule and the Theorem of Total Probability. □

Remark 1.76. As discussed in Remark 1.73, in the statement of Theorem 1.75, we may replace the exhaustiveness of E_i's by the condition $\mathbb{P}(\cup_{i\in\mathcal{I}} E_i) = 1$.

Remark 1.77. In the setup of Theorem 1.72 and Theorem 1.75, information about the 'standard' events E_i's may be known beforehand and we want to understand the probability of occurrence of an arbitrary event E, treated as an 'effect' caused by the E_i's. These two results, therefore, allows us to understand/quantify the relationship between 'cause' and 'effect', in terms of conditional probability.

Notation 1.78. In the setup of the Bayes' Theorem, we shall refer to $\mathbb{P}(E_i), i \in \mathcal{I}$ as prior probabilities and $\mathbb{P}(E_i \mid E), i \in \mathcal{I}$ as posterior probabilities.

Example 1.79. Consider a rare disease X that affects one in a million people. A medical test is used to test for the presence of the disease. The test is 99% accurate in the sense that if a person does not have this disease, the chance that the test shows positive is 1% and if the person has this disease, the chance that the test shows negative is also 1%. Suppose a person is tested for the disease and the test result is positive. Let A be the event that the person has the disease X. Let B be the event that the test shows positive. As per the given information, the given data may be summarized as follows.

$$\mathbb{P}(A) = 10^{-6},\ \mathbb{P}(B^c \mid A) = 0.01,\ \mathbb{P}(B \mid A^c) = 0.01.$$

Then

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) = 1 - 10^{-6},\ \mathbb{P}(B \mid A) = 1 - \mathbb{P}(B^c \mid A) = 0.99.$$

We are interested in the conditional probability that the person has the disease, given that the test result is positive. Here, A and A^c are mutually exclusive and exhaustive. By the Bayes' Theorem, we have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\mathbb{P}(A^c)} = \frac{0.99 \times 10^{-6}}{0.99 \times 10^{-6} + 0.01 \times (1 - 10^{-6})} = 0.000099.$$

Definition 1.80 (Independence of Two events)**.** Let A, B be events in a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Note 1.81. If $A = \emptyset$, then A is independent of any other event B. To see this, observe that

$$\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(A)\mathbb{P}(B) = 0 \times \mathbb{P}(B) = 0.$$

Again, if $A = \Omega$, then A is independent of any other event. Observe that

$$\mathbb{P}(A \cap B) = \mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B).$$

Remark 1.82 (Independence and Mutually Exclusiveness (Pairwise disjointness) of two events). Independence should not be confused with pairwise disjointness! If A and B are disjoint, $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$ and hence A and B can be independent if and only if at least one of $\mathbb{P}(A)$ or $\mathbb{P}(B)$ equals 0. If A and B are disjoint, then $A \subseteq B^c$ and $B \subseteq A^c$. If we know that A has occurred, then we immediately conclude that B did not occur. Independence is not to be expected in such situations. On the other hand, if $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$ and A, B are independent, then $\mathbb{P}(A \cap B) > 0$. In this situation, A and B cannot be mutually exclusive.

Remark 1.83 (Conditional probability and Independence of two events). If A, B are independent and $\mathbb{P}(A) > 0$, then $\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \mathbb{P}(B)$.

Example 1.84. Recall Example 1.66, where we considered rolling a standard six-sided die twice at random. The sample space is $\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$ with $\mathbb{P}((i, j)) = \frac{1}{36}, \forall (i, j) \in \Omega$. Consider the events $A = \{(i, j) : i \text{ is odd}\}$, $B = \{(i, j) : i + j = 4\}$ and $C = \{(i, j) : j = 2\}$.

Observe that

$$\mathbb{P}(A) = \frac{1}{2}, \mathbb{P}(B) = \frac{1}{12}, \mathbb{P}(A \cap B) = \frac{1}{18}.$$

Here, A and B are not independent. Again,

$$\mathbb{P}(A) = \frac{1}{2}, \mathbb{P}(C) = \frac{1}{6}, \mathbb{P}(A \cap C) = \frac{1}{12}.$$

Here, A and C are independent.

Definition 1.85 (Mutual Independence of a collection of events). (a) Let $\{E_1, E_2, \cdots, E_n\}$ be a finite collection of events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that this collection of events is mutually independent or equivalently, the events are mutually independent, if for all $k \in \{2, 3, \cdots, n\}$ and indices $1 \leq i_1 < i_2 < \cdots < i_k \leq n$, we have

$$\mathbb{P}\left(\bigcap_{j=1}^{k} E_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}\left(E_{i_j}\right).$$

(b) Let $\mathcal{I}$ be any indexing set and let $\{E_i : i \in \mathcal{I}\}$ be a collection of events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that this collection of events is mutually independent or equivalently, the events are mutually independent, if all finite subcollections $\{E_{i_i}, E_{i_2}, \cdots, E_{i_k}\}$ are mutually independent.

Definition 1.86 (Pairwise Independence of a collection of events). Let $\mathcal{I}$ be any indexing set and let $\{E_i : i \in \mathcal{I}\}$ be a collection of events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that this collection of events is pairwise independent or equivalently, the events are pairwise independent, if for all distinct indices i and j, the events E_i and E_j are independent.

Note 1.87. To check that events $E_1, E_2, \cdots, E_n$ are mutually independent, we need to check $\binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n} = 2^n - n - 1$ conditions. However, to check that these events are pairwise independent we need to check $\binom{n}{2}$ conditions.

Remark 1.88 (Mutual independence and Pairwise independence of a collection of events). If a collection of events is mutually independent, then by definition the events are also pairwise independent. We consider an example to see that the converse need not be true. Consider a random experiment $\mathcal{E}$

with sample space $\Omega = \{1, 2, 3, 4\}$ and event space $\mathcal{F} = 2^{\Omega}$. If the outcomes are equally likely, then we have the probability function/measure $\mathbb{P}$ determined by the information $\mathbb{P}(\{\omega\}) = \frac{1}{4}, \forall \omega \in \Omega$. Consider the events $A = \{1, 4\}, B = \{2, 4\}, C = \{3, 4\}$. Then $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$. Moreover, $A \cap B = B \cap C = C \cap A = \{4\}$ and hence $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap C) = \mathbb{P}(C \cap A) = \frac{1}{4}$. Therefore, the collection of events $\{A, B, C\}$ is pairwise independent. However, $A \cap B \cap C = \{4\}$ and hence $\mathbb{P}(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$. Here, the collection $\{A, B, C\}$ is not mutually independent.

Notation 1.89. We say that a collection of events is independent or equivalently, some events are independent to mean that the collection is (equivalently, the events are) mutually independent.

Remark 1.90. Consider random experiments where we roll a standard six-sided die thrice or draw two balls from a bin/box/urn. In such experiments multiple draws/trials of the same operations are being performed. In such situations, if the events depending purely on different trials are independent, then we say that the trials are being performed independently. In Example 1.84, the two throws/rolls of the die are independent. Here, the events A and C which depend on the first and the second throws respectively, are independent. If we toss a fair coin twice independently, then the sample space is $\Omega = \{HH, HT, TH, TT\}$ with $\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}) = \frac{1}{4}$.

Note 1.91 (Functions defined on sample spaces)**.** While studying a random phenomena with the framework of a random experiment, in most situations we shall be interested in numerical quantities associated with the outcomes. To elaborate, consider the following two examples.

(a) Consider the random experiment of tossing a coin once. As discussed earlier, the sample space is $\Omega = \{H, T\}$. Suppose we think of the occurrence of head as winning a rupee and the occurrence of a tail as losing a rupee. This information may be captured by a function $X : \Omega \to \mathbb{R}$ given by

$$X(H) := 1, \quad X(T) := -1.$$

(b) Consider the random experiment of tossing a coin until head appears. The sample space may be written as $\Omega = \{H, TH, TTH, TTTH, \cdots\}$. If we are interested in the number of tosses required to obtain the first head, then the information can be captured by the function $X : \Omega \to \mathbb{R}$ given by

$$X(H) := 1, \quad X(TH) := 2, \quad X(TTH) := 3, \quad X(TTTH) := 4, \cdots$$

Now, we focus on analysis of such functions X defined on the sample space Ω of some random experiment $\mathcal{E}$.

Notation 1.92 (Pre-image of a set under a function)**.** Let Ω be a non-empty set and let $X : \Omega \to \mathbb{R}$ be a function. Given any subset A of $\mathbb{R}$, we consider the subset $X^{-1}(A)$ of Ω defined by

$$X^{-1}(A) := \{\omega \in \Omega : X(\omega) \in A\}.$$

The set $X^{-1}(A)$ shall be referred to as the pre-image of A under the function X.

Remark 1.93*.* In Notation 1.92, we do not know whether the function X is bijective. As such, we cannot identify X^{-1} as the 'inverse' function of X. To avoid any confusion, treat $X^{-1}(A)$ as one symbol referring to the set as defined above and do not identify it as a combination of symbols X^{-1} and A.

Remark 1.94 (Shorthand notation for Pre-images)*.* In the setting of Notation 1.92, we shall suppress the symbols ω and use the following notation for convenience, viz.

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} = (X \in A).$$

For specific sets A, other notations, again for convenience, may be used. For example for

(a) If $A = (-\infty, x]$, then we write

$$X^{-1}(A) = (X \in A) = \{\omega \in \Omega : X(\omega) \in (-\infty, x]\} = \{\omega \in \Omega : X(\omega) \leq x\} = (X \leq x).$$

For $A = (-\infty, x), (x, \infty), [x, \infty)$, we shall write $X^{-1}(A)$ to be equal to $(X < x), (X > x), (X \geq x)$ respectively.

(b) If $A = \{x\}$, then we write

$$X^{-1}(A) = (X \in A) = \{\omega \in \Omega : X(\omega) \in \{x\}\} = \{\omega \in \Omega : X(\omega) = x\} = (X = x).$$

Remark 1.95 (Properties of pre-images). Let $X : \Omega \to \mathbb{R}$ be a function. The following are some properties of the pre-images under X, which follow from the fact that X is a function.

(a) $X^{-1}(\mathbb{R}) = \Omega$.

(b) $X^{-1}(\emptyset_{\mathbb{R}}) = \emptyset_{\Omega}$, where $\emptyset_{\mathbb{R}}$ and $\emptyset_{\Omega}$ denote the empty sets under $\mathbb{R}$ and Ω, respectively. When there is no chance of confusion, we simply write $X^{-1}(\emptyset) = \emptyset$.

(c) For any two subsets A, B of $\mathbb{R}$ with $A \cap B = \emptyset$, we have $X^{-1}(A) \cap X^{-1}(B) = \emptyset$.

(d) For any subset A of $\mathbb{R}$, we have $X^{-1}(A^c) = \left(X^{-1}(A)\right)^c$.

(e) Let $\mathcal{I}$ be an indexing set. For any collection $\{A_i : i \in \mathcal{I}\}$ of subsets of $\mathbb{R}$, we have

$$X^{-1}\left(\bigcup_{i\in\mathcal{I}} A_i\right) = \bigcup_{i\in\mathcal{I}} X^{-1}(A_i), \quad X^{-1}\left(\bigcap_{i\in\mathcal{I}} A_i\right) = \bigcap_{i\in\mathcal{I}} X^{-1}(A_i).$$

The above properties shall be used frequently throughout the course.

Note 1.96. As discussed in Note 1.91, we now look at real valued functions defined on Ω, where Ω is the sample space of a random experiment $\mathcal{E}$. We shall also assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. The event space $\mathcal{F}$ shall be taken as the power set 2^{Ω}, unless stated otherwise.

Definition 1.97 (Random variable or RV)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Any real valued function $X : \Omega \to \mathbb{R}$ shall be referred to as a random variable or simply, an RV. In this case, we shall say that X is an RV defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Note 1.98. Since $\mathcal{F}$ is taken to be 2^{Ω}, we immediately have

$$X^{-1}(A) = (X \in A) \in \mathcal{F}$$

for any subset A of $\mathbb{R}$. If $\mathcal{F}$ is taken to be a smaller collection of subsets of Ω, then the above observation may not hold for any arbitrary function X. Given such $\mathcal{F}$, we then restrict our attention to the class of functions X satisfying the above property and refer to them as RVs. It is therefore important to specify $\mathcal{F}$ before we discuss RVs X. As mentioned earlier, $\mathcal{F}$ shall be taken as 2^{Ω}, unless stated otherwise.

Note 1.99. The probability function/measure $\mathbb{P}$ has not been used in the definition of an RV X. We now discuss the role of $\mathbb{P}$ in analysis of RVs X.

Notation 1.100. We write $\mathbb{B}$ to denote the power set of $\mathbb{R}$.

Notation 1.101. Let X be an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then for all $A \in \mathbb{B}$, we have $X^{-1}(A) \in \mathcal{F}$ and hence $\mathbb{P}(X^{-1}(A))$ is well defined. We denote this in terms of a set function $\mathbb{P} \circ X^{-1} : \mathbb{B} \to [0,1]$ given by $\mathbb{P} \circ X^{-1}(A) := \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A), \forall A \in \mathbb{B}$. A shorthand notation $\mathbb{P}_X$ shall also be used to refer to $\mathbb{P} \circ X^{-1}$.

Notation 1.102. Similar to the discussion in Remark 1.94, we shall write $\mathbb{P}(X \leq x), \mathbb{P}(X = x)$ etc. for $\mathbb{P} \circ X^{-1}(A)$ where $A = (-\infty, x], \{x\}$ etc. respectively.

Proposition 1.103. *Let X be an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, the set function $\mathbb{P} \circ X^{-1}$ is a probability function/measure defined on the collection $\mathbb{B}$.*

Proof. We verify the axioms/properties of a probability function/measure as mentioned in Definition 1.33.

We have $\mathbb{P} \circ X^{-1}(\mathbb{R}) = \mathbb{P}(X^{-1}(\mathbb{R})) = \mathbb{P}(\Omega) = 1$. Since $\mathbb{P}$ is a probability measure on $\mathcal{F}$, we also have $\mathbb{P} \circ X^{-1}(A) = \mathbb{P}(X^{-1}(A)) \geq 0, \forall A \in \mathbb{B}$.

If $\{A_n\}_n$ is a sequence of pairwise disjoint sets in $\mathbb{B}$, then $\{X^{-1}(A_n)\}_n$ is a sequence of pairwise disjoint events in $\mathcal{F}$. Hence,

$$\mathbb{P} \circ X^{-1}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} X^{-1}(A_n)\right) = \sum_{n=1}^{\infty} \mathbb{P}(X^{-1}(A_n)) = \sum_{n=1}^{\infty} \mathbb{P} \circ X^{-1}(A_n).$$

This proves countable additivity property for $\mathbb{P} \circ X^{-1}$ and the proof is complete. □

Definition 1.104 (Induced Probability Space and Induced Probability Measure)**.** If X is an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the probability function/measure $\mathbb{P} \circ X^{-1}$ on $\mathbb{B}$ is referred to as the induced probability function/measure induced by X. In this case, $(\mathbb{R}, \mathbb{B}, \mathbb{P} \circ X^{-1})$ is referred to as the induced probability space induced by X.

Example 1.105. Recall from Remark 1.90, that if we toss a fair coin twice independently, then the sample space is $\Omega = \{HH, HT, TH, TT\}$ with $\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}) = \frac{1}{4}$. Consider the RV $X : \Omega \to \mathbb{R}$ which denotes the number of heads. Here,

$$X(HH) = 2, \quad X(HT) = X(TH) = 1, \quad X(TT) = 0.$$

Consider the induced probability measure $\mathbb{P} \circ X^{-1}$ on $\mathbb{B}$. We have

$$\mathbb{P} \circ X^{-1}(\{0\}) = \mathbb{P}(X^{-1}(\{0\})) = \mathbb{P}(\{TT\}) = \frac{1}{4},$$

$$\mathbb{P} \circ X^{-1}(\{1\}) = \mathbb{P}(X^{-1}(\{1\})) = \mathbb{P}(\{HT, TH\}) = \frac{1}{2},$$

$$\mathbb{P} \circ X^{-1}(\{2\}) = \mathbb{P}(X^{-1}(\{2\})) = \mathbb{P}(\{HH\}) = \frac{1}{4}.$$

More generally, for any $A \in \mathbb{B}$, we have

$$\mathbb{P} \circ X^{-1}(A) = \mathbb{P}(\{\omega : X(\omega) \in A\}) = \sum_{i \in \{0,1,2\} \cap A} \mathbb{P} \circ X^{-1}(\{i\}).$$

Remark 1.106. If we know the probability function/measure $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ for any RV X, then we get the information about all the probabilities $\mathbb{P}(X \in A), A \in \mathbb{B}$ for events $X^{-1}(A) = (X \in A), A \in \mathbb{B}$ involving the RV X. In what follows, our analysis of RV X shall be through the understanding of probability function/measure $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ on $\mathbb{B}$.

Definition 1.107 (Law/Distribution of an RV)**.** If X is an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the probability function/measure $\mathbb{P}_X$ on $\mathbb{B}$ is referred to as the law or distribution of the RV X.

We now discuss some properties of a probability function/measure. To do this, we first introduce a concept involving sequence of sets.

Definition 1.108 (Increasing and decreasing sequence of sets)**.** Let $\{A_n\}_n$ be a sequence of subsets of a non-empty set Ω.

(a) If $A_n \subseteq A_{n+1}, \forall n = 1, 2, \cdots$, we say that the sequence $\{A_n\}_n$ is increasing. In this case, we say A_n increases to A, denoted by $A_n \uparrow A$, where $A = \bigcup_{n=1}^{\infty} A_n$.

(b) If $A_n \supseteq A_{n+1}, \forall n = 1, 2, \cdots$, we say that the sequence $\{A_n\}_n$ is decreasing. In this case, we say A_n decreases to A, denoted by $A_n \downarrow A$, where $A = \bigcap_{n=1}^{\infty} A_n$.

Remark 1.109. $A_n \uparrow A$ if and only if $A_n^c \downarrow A^c$.

Proposition 1.110 (Continuity of a probability measure)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space.*

(a) (Continuity from below) Let $\{A_n\}_n$ be sequence in $\mathcal{F}$, such that $A_n \uparrow A$. Then

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A).$$

(b) (Continuity from above) Let $\{A_n\}_n$ be sequence in $\mathcal{F}$, such that $A_n \downarrow A$. Then

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A).$$

Proof. To prove the first statement. Since $\{A_n\}_n$ is an increasing sequence of sets, we have

$$A_n \cap (A_1 \cup A_2 \cup \cdots \cup A_{n-1})^c = A_n \cap A_{n-1}^c, \forall n \geq 2.$$

Then using a hint from practice problem set 1, we have

$$\bigcup_{n=1}^{\infty} A_n = A_1 \cup \left(\bigcup_{n=2}^{\infty} (A_n \cap A_{n-1}^c) \right).$$

Since the sets $A_1, A_2 \cap A_1^c, A_3 \cap A_2^c, \cdots$ are pairwise disjoint, using the countable additivity of $\mathbb{P}$, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) &= \mathbb{P}(A_1) + \sum_{n=2}^{\infty} \mathbb{P}(A_n \cap A_{n-1}^c) = \mathbb{P}(A_1) + \lim_{k\to\infty} \sum_{n=2}^{k} \mathbb{P}(A_n \cap A_{n-1}^c) \\ &= \mathbb{P}(A_1) + \lim_{k\to\infty} \sum_{n=2}^{k} [\mathbb{P}(A_n) - \mathbb{P}(A_{n-1})] \\ &= \mathbb{P}(A_1) + \lim_{k\to\infty} [\mathbb{P}(A_k) - \mathbb{P}(A_1)] = \lim_{k\to\infty} \mathbb{P}(A_k).\end{aligned}$$

This completes the proof of the first statement.

To prove the second statement. First observe that $A_n^c \uparrow A^c$ with $A = \bigcap_{n=1}^{\infty} A_n$. Using the first statement, we have

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \lim_{n\to\infty} \mathbb{P}(A_n^c) = \lim_{n\to\infty} [1 - \mathbb{P}(A_n^c)] = \lim_{n\to\infty} \mathbb{P}(A_n).$$

The proof is complete. □

Definition 1.111 (Distribution function of an RV)**.** Let X be an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with law/distribution $\mathbb{P}_X$. Consider the function $F_X : \mathbb{R} \to \mathbb{R}$ defined by $F_X(x) :=$

$\mathbb{P}_X((-\infty, x]) = \mathbb{P}(X \leq x), \forall x \in \mathbb{R}$. The function F_X is called the cumulative distribution function (CDF) or simply, the distribution function (DF) of the RV X.

Remark 1.112 (RVs equal in law/distribution). Let X be an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let Y be an RV defined on a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$. If $\mathbb{P} \circ X^{-1} = \mathbb{P}' \circ Y^{-1}$, i.e. $\mathbb{P} \circ X^{-1}(A) = \mathbb{P}' \circ Y^{-1}(A), \forall A \in \mathbb{B}$, then we say that X and Y are equal in law/distribution. In this case, $F_X = F_Y$, i.e. $F_X(x) = F_Y(x), \forall x \in \mathbb{R}$.

Remark 1.113. Let X and Y be two RVs, possibly defined on different probability spaces. If $F_X = F_Y$, then it can be shown that X and Y are equal in law/distribution. The proof of this statement is beyond the scope of this course. This statement is often restated as 'the DF of an RV uniquely determines the law/distribution of the RV'.

Example 1.114. Consider X as in Example 1.105. Then for all $x \in \mathbb{R}$, we have

$$F_X(x) = \mathbb{P}_X((-\infty, x])) = \sum_{i \in \{0,1,2\} \cap (-\infty, x]} \mathbb{P}_X(\{i\}) = \begin{cases} 0, \text{ if } x < 0, \\ \mathbb{P}_X(\{0\}), \text{ if } 0 \leq x < 1, \\ \mathbb{P}_X(\{0\}) + \mathbb{P}_X(\{1\}), \text{ if } 1 \leq x < 2, \\ \mathbb{P}_X(\{0\}) + \mathbb{P}_X(\{1\}) + \mathbb{P}_X(\{2\}), \text{ if } x \geq 2. \end{cases}$$

Therefore,

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

Theorem 1.115. *Let X be an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with law $\mathbb{P}_X$ and DF F_X. Then*

(a) F_X is non-decreasing, i.e. $F_X(x) \leq F_Y(y), \forall x < y$.

(b) F_X is right continuous, i.e. $\lim_{h \downarrow 0} F_X(x+h) = F_X(x), \forall x \in \mathbb{R}$.

(c) $F_X(-\infty) := \lim_{x \to -\infty} F_X(x) = 0$ and $F_X(\infty) := \lim_{x \to \infty} F_X(x) = 1$.

Proof. For all $x < y$, observe that $(-\infty, x] \subsetneq (-\infty, y]$. Since $\mathbb{P}_X$ is a probability measure, we have $\mathbb{P}_X((-\infty, x]) \leq \mathbb{P}_X((-\infty, y])$. The statement (a) follows.

By definition, F_X takes values in $[0, 1]$ and hence it is bounded. Since F_X is non-decreasing, the limit $F_X(x+) = \lim_{h \downarrow 0} F_X(x + h)$ exists for all $x \in \mathbb{R}$. Using the non-decreasing property, we use the following fact from real analysis that $F_X(x+) = \lim_{n\to\infty} F_X(x + \frac{1}{n})$. By Proposition 1.110, we have

$$F_X(x+) = \lim_{n\to\infty} F_X\left(x + \frac{1}{n}\right) = \lim_{n\to\infty} \mathbb{P}_X(\left(-\infty, x + \frac{1}{n}\right]) = \mathbb{P}_X((-\infty, x]) = F_X(x).$$

This proves statement (b). Here, we use the fact that $(-\infty, x + \frac{1}{n}] \downarrow (-\infty, x]$.

Similar to the proof of statement (b), we have

$$F_X(-\infty) = \lim_{n\to\infty} F_X(-n) = \lim_{n\to\infty} \mathbb{P}_X((-\infty, -n]) = \mathbb{P}_X(\emptyset) = 0,$$

and

$$F_X(\infty) = \lim_{n\to\infty} F_X(n) = \lim_{n\to\infty} \mathbb{P}_X((-\infty, n]) = \mathbb{P}_X(\mathbb{R}) = 1.$$

Here, we use that facts that $(-\infty, -n] \downarrow \emptyset$ and $(-\infty, n] \uparrow \mathbb{R}$. This proves statement (c). □

The next theorem is stated without proof. The arguments required to prove this statement is beyond the scope of this course.

Theorem 1.116. *Let $F : \mathbb{R} \to \mathbb{R}$ be a non-decreasing and right continuous function such that $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Then there exists an RV X defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $F = F_X$, i.e. $F(x) = F_X(x), \forall x$.*

Remark 1.117. Given any function $F : \mathbb{R} \to \mathbb{R}$, as soon as we check the relevant conditions, we can claim that it is the DF of some RV by Theorem 1.116.

Example 1.118. Consider the function $F : \mathbb{R} \to \mathbb{R}$ defined by

$$F(x) = \begin{cases} 0, \text{ if } x < 0, \\ x, \text{ if } 0 \leq x < 1, \\ 1, \text{ if } x \geq 1. \end{cases}$$

The function is a constant on $(-\infty, 0)$ and on $[1, \infty)$. Moreover, it is non-decreasing in the interval $[0, 1)$. Further for $x < 0, y \in (0, 1), z > 1$, we have

$$F(x) = F(0) < F(y) < F(1) = F(z).$$

Hence, F in non-decreasing over $\mathbb{R}$. Again, by definition F is continuous on the intervals $(-\infty, 0)$, $(0, 1)$ and $(1, \infty)$. We check for right continuity at the points 0 and 1. We have

$$\lim_{h \downarrow 0} F(0 + h) = \lim_{h \downarrow 0} h = 0 = F(0), \quad \lim_{h \downarrow 0} F(1 + h) = \lim_{h \downarrow 0} 1 = 1 = F(1).$$

Hence, F is right continuous on $\mathbb{R}$. Finally, $\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} 0 = 0$ and $\lim_{x \to \infty} F(x) = \lim_{x \to \infty} 1 = 1$. Hence, F is the DF of some RV. Later on, we shall identify the corresponding RV.

Proposition 1.119 (Further properties of a DF)**.** *Let X be an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with law $\mathbb{P}_X$ and DF F_X.*

(a) For all $x \in \mathbb{R}$, the limit $F_X(x-) = \lim_{h \downarrow 0} F_X(x - h)$ exists and equals $\mathbb{P}_X((-\infty, x)) = \mathbb{P}(X < x)$.

Proof. Since F_X in non-decreasing and bounded, as argued in Theorem 1.115, the limit $F_X(x-) = \lim_{h \downarrow 0} F_X(x - h)$ exists and moreover, by Proposition 1.110 we have

$$F_X(x-) = \lim_{n \to \infty} F_X\left(x - \frac{1}{n}\right) = \lim_{n \to \infty} \mathbb{P}_X(\left(-\infty, x - \frac{1}{n}\right]) = \mathbb{P}_X((-\infty, x)) = \mathbb{P}(X < x).$$

Here, we use the fact that $(-\infty, x - \frac{1}{n}] \uparrow (-\infty, x)$. □

(b) For all $x \in \mathbb{R}$, $\mathbb{P}(X \geq x) = 1 - F_X(x-)$.

Proof. We have, $\mathbb{P}(X \geq x) = \mathbb{P}_X([x, \infty)) = \mathbb{P}_X((-\infty, x)^c) = 1 - \mathbb{P}_X((-\infty, x)) = 1 - F_X(x-)$. □

(c) For any $x \in \mathbb{R}$, $F_X(x-) \leq F_X(x+)$.

Proof. By the non-decreasing property of F_X, for all $x \in \mathbb{R}$ and positive integers n, we have, $F_X(x - \frac{1}{n}) \leq F_X(x + \frac{1}{n})$. Letting n go to infinity in this inequality, we get the result. □

(d) F_X is continuous at x if and only if $F_X(x) = F_X(x-)$.

Proof. A real valued function is continuous at a point x if and only if the function is both right continuous and left continuous at the point x. Now, by construction, F_X is right continuous on $\mathbb{R}$. Hence, F_X is continuous at x if and only if F_X is left continuous at x. The last statement is exactly the statement to be proved. □

(e) *Only possible discontinuities of F_X are jump discontinuities.*

Proof. As discussed in Theorem 1.115 and in part (a), for any $x \in \mathbb{R}$, both the limits $F_X(x+)$ and $F_X(x-)$ exist and $F_X(x+) = F_X(x)$. Since $F_X(x-) \leq F_X(x+)$, the only possible discontinuity appears if and only if $F_X(x-) < F_X(x+)$. These discontinuities are jump discontinuities. This completes the proof. □

(f) *For all $x \in \mathbb{R}$, we have $F_X(x+) - F_X(x-) = \mathbb{P}(X = x)$.*

Proof. By the finite additivity of $\mathbb{P}_X$, we have $F_X(x+) - F_X(x-) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = \mathbb{P}_X((-\infty, x]) - \mathbb{P}_X((-\infty, x)) = \mathbb{P}_X(\{x\}) = \mathbb{P}(X = x)$. □

(g) *If F_X has a jump at x, then the jump is given by $F_X(x+) - F_X(x-) = \mathbb{P}(X = x)$.*

Proof. If F_X has a jump at x, then the jump is given by $F_X(x+) - F_X(x-)$. The result follows from statement (f). □

(h) *F_X is continuous at x if and only if $\mathbb{P}(X = x) = 0$.*

Proof. Recall that $F_X(x+) = F_X(x)$. Then by statement (d) and (f), we have F_X is continuous at x if and only if $F_X(x+) = F_X(x-)$ and hence, if and only if $\mathbb{P}(X = x) = 0$. □

(i) *Consider the set $D := \{x \in \mathbb{R} : F_X \text{ is discontinuous at } x\} = \{x \in \mathbb{R} : F_X(x-) < F_X(x+)\} = \{x \in \mathbb{R} : \mathbb{P}(X = x) > 0\}$. Then D is either finite or countably infinite. (Note that if F_X is continuous on $\mathbb{R}$, then $D = \emptyset$.)*

Proof. Left as an exercise in practice problem set 3. □

(j) *For all $x < y$, we have*

$$\mathbb{P}(x < X \leq y) = F_X(y) - F_X(x),$$

$$\mathbb{P}(x < X < y) = F_X(y-) - F_X(x),$$

$$\mathbb{P}(x \leq X < y) = F_X(y-) - F_X(x-),$$

$$\mathbb{P}(x \leq X \leq y) = F_X(y) - F_X(x-).$$

Proof. We prove the first two equalities. Proof of the last two equalities are similar.

By the finite additivity of $\mathbb{P}_X$, we have $F_X(y) - F_X(x) = \mathbb{P}_X((-\infty, y]) - \mathbb{P}_X((-\infty, x]) = \mathbb{P}_X((x, y]) = \mathbb{P}(x < X \leq y)$.

Again, $F_X(y-) - F_X(x) = \mathbb{P}_X((-\infty, y)) - \mathbb{P}_X((-\infty, x]) = \mathbb{P}_X((x, y)) = \mathbb{P}(x < X < y)$.

This completes the proof. □

Example 1.120. Consider the function $F : \mathbb{R} \to \mathbb{R}$ defined by

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

Assume that F is the DF of some RV X (left as an exercise in practice problem set 3). Since F is continuous on the intervals $(-\infty, 0), (0, 1), (1, 2)$ and $(2, \infty)$, discontinuities may arise only at the points $0, 1, 2$.

We have $F(0-) = \lim_{h \downarrow 0} F(0 - h) = 0$ and $F(0) = \frac{1}{4}$. Therefore F is discontinuous at 0 with jump $F(0) - F(0-) = \frac{1}{4}$.

We have $F(1-) = \lim_{h \downarrow 0} F(1 - h) = \lim_{h \downarrow 0}[\frac{1}{4} + \frac{1-h}{2}] = \frac{3}{4}$ and $F(1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$. Therefore F is continuous at 1.

We have $F(2-) = \lim_{h \downarrow 0} F(2 - h) = \lim_{h \downarrow 0}[\frac{1}{2} + \frac{2-h}{4}] = 1$ and $F(2) = 1$. Therefore F is continuous at 2.

Only discontinuity of F is at the point 0. In particular, $\mathbb{P}(X = 0) = F(0) - F(0-) = \frac{1}{4}$. At all other points F is continuous and hence $\mathbb{P}(X = x) = 0, \forall x \neq 0$.

Observe that $\mathbb{P}(0 \leq X < 1) = F(1-) - F(0-) = \frac{3}{4}$. Again, $\mathbb{P}(\frac{3}{2} < X \leq 2) = F(2) - F(\frac{3}{2}) = 1 - [\frac{1}{2} + \frac{3}{8}] = \frac{1}{8}$.

We now discuss special classes of RVs defined on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Recall that $\mathbb{P}_X$ and F_X denote the law/distribution and the distribution function (DF) of an RV X, respectively.

Definition 1.121 (Discrete RV)**.** An RV X is said to be a discrete RV if there exists a finite or countably infinite set $S \subsetneq \mathbb{R}$ such that

$$1 = \mathbb{P}_X(S) = \mathbb{P}(X \in S) = \sum_{x \in S} \mathbb{P}_X(\{x\}) = \sum_{x \in S} \mathbb{P}(X = x)$$

and $\mathbb{P}(X = x) > 0, \forall x \in S$. In this situation, we refer to the set S as the support of the discrete RV X.

Remark 1.122. Let X be a discrete RV with DF F_X and support S. Then we have the following observations.

(a) $\mathbb{P}_X(S^c) = 1 - \mathbb{P}_X(S) = 0$. In particular, for any $x \in S^c$, $0 \leq \mathbb{P}(X = x) = \mathbb{P}_X(\{x\}) \leq \mathbb{P}_X(S^c) = 0$ and hence $\mathbb{P}(X = x) = 0, \forall x \in S^c$.

(b) Since $\mathbb{P}_X(S) = 1$, for any $A \subseteq \mathbb{R}$, we have $\mathbb{P}_X(A) = \mathbb{P}_X(A \cap S)$ (see problem set 1). Moreover,

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}_X(A \cap S) = \sum_{x \in A \cap S} \mathbb{P}(X = x).$$

(c) Recall that F_X is right continuous, i.e. $F_X(x+) = F_X(x), \forall x \in \mathbb{R}$. Moreover, $F_X(x) - F_X(x-) = \mathbb{P}(X = x)$. From the discussion above, we conclude that

$$F_X(x) - F_X(x-) = \mathbb{P}(X = x) \begin{cases} > 0, & \text{if } x \in S, \\ = 0, & \text{if } x \in S^c. \end{cases}$$

Hence, the set of discontinuities of F_X is exactly the support S.

(d) Note that

$$1 = \sum_{x \in S} \mathbb{P}(X = x) = \sum_{x \in S} [F_X(x) - F_X(x-)].$$

Hence, the sum of the jumps of F_X is exactly 1.

Example 1.123. Consider the DF $F : \mathbb{R} \to [0,1]$ considered in Example 1.120 given by

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

As discussed earlier, F only has a discontinuity at the point 0. If an RV X has this F as the DF, then

$$\sum_{x \in D} \mathbb{P}(X = x) = \mathbb{P}(X = 0) = \frac{1}{4} \neq 1,$$

with $D = \{0\}$ as the set of discontinuities of F. This RV X is not discrete.

Example 1.124. Let X denote the number of heads in tossing a fair coin twice independently. As computed earlier in Example 1.114, the DF F_X is given by

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

Clearly, the set D of discontinuities of F_X is $\{0, 1, 2\}$ with

$$\mathbb{P}(X = x) = F_X(x) - F_X(x-) = \begin{cases} \frac{1}{4} - 0 = \frac{1}{4}, \text{ if } x = 0, \\ \frac{3}{4} - \frac{1}{4} = \frac{1}{2}, \text{ if } x = 1, \\ 1 - \frac{3}{4} = \frac{1}{4}, \text{ if } x = 2. \end{cases}$$

Since $\sum_{x \in D} \mathbb{P}(X = x) = 1$, the RV X is discrete with support D.

Definition 1.125 (Probability Mass Function (p.m.f.))**.** Let X be a discrete RV with DF F_X and support S. Consider the function $f_X : \mathbb{R} \to \mathbb{R}$ defined by

$$f_X(x) := \begin{cases} F_X(x) - F_X(x-) = \mathbb{P}(X = x), \text{ if } x \in S, \\ 0, \text{ if } x \in S^c. \end{cases}$$

This function f_X is called the probability mass function (p.m.f.) of X.

Example 1.126. Continuing with the Example 1.124, the p.m.f. f_X is given by

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

Remark 1.127. Let X be a discrete RV with DF F_X, p.m.f. f_X and support S. Then we have the following observations.

(a) Continuing the discussion from Remark 1.122, we have for all $A \subseteq \mathbb{R}$,

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \sum_{x \in A \cap S} f_X(x).$$

(b) As a special case of the previous observation, note that for $A = (-\infty, x], x \in \mathbb{R}$, we obtain

$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(X \in (-\infty, x]) = \sum_{t \in (-\infty, x] \cap S} f_X(t).$$

Therefore, the p.m.f. f_X is uniquely determined by the DF F_X and vice versa.

(c) To study a discrete RV X, we may study any one of the following three quantities, viz. the law/distribution $\mathbb{P}_X$, the DF F_X or the p.m.f. f_X. Given any one of these quantities, the other two can be obtained using the relations described above.

(d) By Definition 1.121 and Definition 1.125, we have that the p.m.f. $f_X : \mathbb{R} \to \mathbb{R}$ is a function such that

$$f_X(x) = 0, \forall x \in S^c, \quad f_X(x) > 0, \forall x \in S, \quad \sum_{x \in S} f_X(x) = 1.$$

Remark 1.128. Let $\emptyset \neq S \subset \mathbb{R}$ be a finite or countably infinite set and let $f : \mathbb{R} \to \mathbb{R}$ be such that

$$f(x) = 0, \forall x \in S^c, \quad f(x) > 0, \forall x \in S, \quad \sum_{x \in S} f(x) = 1.$$

Then by an argument similar to Proposition 1.45, we conclude that $\mathbb{P}$ as defined below is a probability function/measure on $\mathbb{B}$, where $\mathbb{B}$ denotes the power set of $\mathbb{R}$. For all $A \subseteq \mathbb{R}$, consider

$$\mathbb{P}(A) := \sum_{x \in A \cap S} f(x).$$

By an argument similar to Theorem 1.115, we can then show that the function $F : \mathbb{R} \to \mathbb{R}$ defined by $F(x) := \mathbb{P}((-\infty, x]), \forall x \in \mathbb{R}$ is non-decreasing, right continuous with $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. By Theorem 1.116, this F is the DF of some RV Y, i.e. $F_Y = F$ and by construction, Y must be discrete with support S and p.m.f. $f_Y = f$.

Example 1.129. Take S to be the set of natural numbers $\{1, 2, \cdots\}$ and consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := \begin{cases} \frac{1}{2^x}, \text{ if } x \in S, \\ 0, \text{ if } x \in S^c. \end{cases}$$

Then f takes non-negative values with $\sum_{x \in S} f(x) = \sum_{n=1}^{\infty} \frac{1}{2^n} = 1$. Therefore f is the p.m.f. of some RV X with DF F_X given by

$$\begin{aligned} F_X(x) = \mathbb{P}(X \leq x) &= \sum_{t \in (-\infty, x] \cap S} f_X(t) \\ &= \begin{cases} 0, \text{ if } x < 1, \\ \sum_{n=1}^{m} \frac{1}{2^n}, \text{ if } x \in [m, m+1), m \in S. \end{cases} = \begin{cases} 0, \text{ if } x < 1, \\ 1 - \frac{1}{2^m}, \text{ if } x \in [m, m+1), m \in S. \end{cases} \end{aligned}$$

Definition 1.130 (Continuous RV and its Probability Density Function (p.d.f.))**.** An RV X is said to be a continuous RV if there exists an integrable function $f : \mathbb{R} \to [0, \infty)$ such that

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)\, dt, \forall x \in \mathbb{R}.$$

The function f is called the probability density function (p.d.f.) of X.

Remark 1.131. Let X be a continuous RV with DF F_X and p.d.f. f_X. Then we have the following observations.

(a) Since f_X is integrable, from the relation $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \forall x \in \mathbb{R}$, we have F_X is continuous on $\mathbb{R}$. In particular, F_X is absolutely continuous. Moreover, for all $a < b$, we have

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

(b) Since F_X is continuous, we have

(i) $F_X(x-) = F_X(x) = F_X(x+), \forall x \in \mathbb{R}$.

(ii) $\mathbb{P}(X = x) = \mathbb{P}_X(\{x\}) = F_X(x) - F_X(x-) = 0, \forall x \in \mathbb{R}$.

(iii) $\mathbb{P}(X < x) = F_X(x-) = F_X(x) = \mathbb{P}(X \leq x), \forall x \in \mathbb{R}$.

(iv) For all $a < b$,

$$\begin{aligned}\mathbb{P}(a < X < b) &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a \leq X \leq b)\\ &= F_X(b) - F_X(a) = \int_{a}^{b} f_X(t)\, dt.\end{aligned}$$

(c) If $A \subset \mathbb{R}$ is finite or countably infinite, then by the finite/countable additivity of $\mathbb{P}_X$, we have

$$\mathbb{P}(X \in A) = \mathbb{P}_X(A) = \sum_{x \in A} \mathbb{P}_X(\{x\}) = 0.$$

(d) By definition, we have $f_X(x) \geq 0, \forall x \in \mathbb{R}$ and

$$1 = \lim_{x \to \infty} F_X(x) = \lim_{x \to \infty} \int_{-\infty}^{x} f_X(t)\, dt = \int_{-\infty}^{\infty} f_X(t)\, dt.$$

Remark 1.132. Let $f : \mathbb{R} \to [0, \infty)$ be an integrable function with $\int_{-\infty}^{\infty} f(t)\, dt = 1$. Then the function $F : \mathbb{R} \to [0, 1]$ defined by $F(x) := \int_{-\infty}^{x} f(t)\, dt, \forall x \in \mathbb{R}$ is non-decreasing and continuous with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. By Theorem 1.116, this F is the DF of some RV Y, i.e. $F_Y = F$ and by construction, Y must be continuous with p.d.f. $f_Y = f$.

Example 1.133. Let X be an RV with the DF $F_X : \mathbb{R} \to \mathbb{R}$ as discussed in Example 1.118. Here,

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

Then the function $f : \mathbb{R} \to [0, \infty)$ defined by

$$f(x) := \begin{cases} 1, \text{ if } 0 \leq x \leq 1, \\ 0, \text{ otherwise} \end{cases}$$

is an integrable function with $F_X(x) = \int_{-\infty}^{x} f(t)\, dt, \forall x \in \mathbb{R}$. Therefore, X is a continuous RV with p.d.f. f.

Example 1.134. Consider the DF $F : \mathbb{R} \to [0, 1]$ considered in Example 1.120 given by

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

As discussed earlier, F has a discontinuity at the point 0. Therefore, an RV X with DF F is not a continuous RV.

Note 1.135. Given a continuous RV X with p.d.f. f_X, the DF F_X is computed by the formula $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \forall x \in \mathbb{R}$.

Example 1.136. Consider a function $f : \mathbb{R} \to \mathbb{R}$ of the form

$$f(x) = \begin{cases} \alpha x, & \text{if } x \in [-1, 0), \\ \frac{x^2}{8}, & \text{if } x \in [0, 2], \\ 0, & \text{otherwise} \end{cases}$$

for some $\alpha \in \mathbb{R}$. For this f to be a p.d.f. of a continuous RV, two conditions need to be satisfied, viz. $f(x) \geq 0, \forall x \in \mathbb{R}$ and $\int_{-\infty}^{\infty} f(x)\,dx = 1$.
The first condition is satisfied on $(-\infty, -1) \cup [0, \infty)$. For $x \in [-1, 0)$, we must have $\alpha x \geq 0$, which implies $\alpha \leq 0$.
From the second condition, we have $\int_{-1}^{0} \alpha x\,dx + \int_{0}^{2} \frac{x^2}{8}\,dx = 1$. This yields $\alpha = -\frac{4}{3}$, which satisfies $\alpha \leq 0$.
Therefore, for f to be a p.d.f. we must have $\alpha = -\frac{4}{3}$.

In what follows, we consider the question of computing f_X from the DF F_X.

Remark 1.137 (Is the p.d.f. of a continuous RV unique?). Let X be a continuous RV with DF F_X and p.d.f. f_X. Fix any finite or countably infinite set $A \subset \mathbb{R}$ and fix $c \geq 0$. Consider the function $g : \mathbb{R} \to [0, \infty)$ defined by

$$g(x) := \begin{cases} f_X(x), & \text{if } x \in A^c, \\ c, & \text{if } x \in A. \end{cases}$$

Then g is integrable and $F_X(x) = \int_{-\infty}^{x} g(t)\,dt, \forall x \in \mathbb{R}$. Hence, g is also a p.d.f. for X. Therefore, the RV X with DF F_X is a continuous RV with p.d.f. f (or g). For example,

$$g(x) := \begin{cases} 1, & \text{if } 0 < x < 1, \\ 0, & \text{otherwise} \end{cases}$$

is a p.d.f. for X as in Example 1.133. More generally, we may also consider

$$g(x) := \begin{cases} f_X(x), & \text{if } x \in A^c, \\ c_x, & \text{if } x \in A \end{cases}$$

as a p.d.f., where $c_x \geq 0, \forall x \in A$.

Note 1.138. In fact, a p.d.f. f_X for a continuous RV X is determined uniquely on the complement of sets of 'length 0', such as sets which are finite or countably infinite. We do not make a precise statement – this is beyond the scope of this course. However, we consider the deduction of p.d.f.s from the DFs.

The next result is stated without proof.

Theorem 1.139. *Let X be an RV with DF F_X.*

(a) If F_X is differentiable on $\mathbb{R}$ with $\int_{-\infty}^{\infty} F_X'(t)\,dt = 1$, then X is a continuous RV with p.d.f. F_X'.

(b) If F_X is differentiable everywhere except on a finite or a countably infinite set $A \subset \mathbb{R}$ with $\int_{-\infty}^{\infty} F_X'(t)\,dt = 1$, then X is a continuous RV with p.d.f. f given by

$$f(x) := \begin{cases} F_X'(x), & \textit{if } x \in A^c, \\ 0, & \textit{if } x \in A. \end{cases}$$

Note 1.140. Continuing the discussion from Note 1.135, the DF F_X of a continuous RV X may be used to compute the p.d.f. f_X. In Example 1.133, the DF F_X is given by

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

It is differentiable everywhere except at the points 0 and 1. Using Theorem 1.139, we have the p.d.f. given by

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

Note 1.141. To study a continuous RV X, we may study any one of the following three quantities, viz. the law/distribution $\mathbb{P}_X$, the DF F_X or the p.d.f. f_X. Given any one of these quantities, the other two can be obtained using the relations described above.

Definition 1.142 (Support of a Continuous RV)**.** Let X be a continuous RV with DF F_X. The set

$$S := \{x \in \mathbb{R} : F_X(x+h) - F_X(x-h) > 0, \forall h > 0\}$$

is defined to be the support of X.

Remark 1.143*.* The support S of a continuous RV X can be expressed in terms of the law/distribution of X as follows.

$$S = \{x \in \mathbb{R} : \mathbb{P}(x-h < X \leq x+h) > 0, \forall h > 0\} = \{x \in \mathbb{R} : \mathbb{P}_X((x-h, x+h]) > 0, \forall h > 0\}.$$

Remark 1.144*.* The support S of a continuous RV X can be expressed in terms of the p.d.f. f_X as follows.

$$S = \{x \in \mathbb{R} : \int_{x-h}^{x+h} f_X(t)\, dt > 0, \forall h > 0\}.$$

Note 1.145. If $x \notin S$, where S is the support of a continuous RV X, then there exists $h > 0$ such that $F_X(x+h) = F_X(x-h)$. By the non-decreasing property of F_X, we conclude that F_X remains a constant on the interval $[x-h, x+h]$. In particular, $f_X(t) = F_X'(t) = 0, \forall t \in (x-h, x+h)$.

Example 1.146. Consider a continuous RV X with DF $F_X : \mathbb{R} \to [0,1]$ and p.d.f. $f_X : \mathbb{R} \to [0, \infty)$ given by

$$F_X(x) := \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } 0 \leq x < 1, \\ 1, & \text{if } x \geq 1. \end{cases}, \qquad f_X(x) := \begin{cases} 1, & \text{if } 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

To identify the support S, we consider the following cases.

(a) Let $x \in (-\infty, 0)$. Then for all h with $-x > h > 0$, we have $x - h < x + h < 0$ and consequently, $F_X(x+h) - F_X(x-h) = 0 - 0 = 0$. Therefore $x \notin S$.

(b) Let $x \in (1, \infty)$. Then for all $0 < h < x - 1$, we have $1 < x - h < x + h$ and consequently, $F_X(x+h) - F_X(x-h) = 1 - 1 = 0$. Therefore $x \notin S$.

(c) Let $x \in (0,1)$. For any $0 < h < \min\{x, 1-x\}$, we have $0 < x - h < x + h < 1$ and consequently, $F_X(x+h) - F_X(x-h) = (x+h) - (x-h) = 2h > 0$. For $h \geq \min\{x, 1-x\}$, at least one of $x-h, x+h$ is in $(0,1)^c$ and hence $F_X(x+h) - F_X(x-h) > 0$. Therefore $x \in S$.

(d) Let $x = 0$. Then for any $h > 0$, we have $F_X(0+h) - F_X(0-h) = F_X(0+h) > 0$. Then $0 \in S$. By a similar argument, $1 \in S$.

From the above discussion, we conclude that $S = [0,1]$.

Remark 1.147 (Identifying discrete/continuous RVs from their DFs)*.* Suppose that the distribution of an RV X is specified by a given DF F_X. In order to check if X is a discrete/continuous RV, we use the following steps.

(a) Identify the set $D = \{x \in \mathbb{R} : F_X(x-) < F_X(x+)\} = \{x \in \mathbb{R} : \mathbb{P}(X = x) > 0\}$ of discontinuities of F_X. Recall that D is a finite or a countably infinite set.

(b) If D is empty, then F_X is continuous on $\mathbb{R}$. By verifying the hypothesis of Theorem 1.139 or otherwise, check if there exists a p.d.f.. If a p.d.f. exists, then X is a continuous RV. Otherwise, X is not a continuous RV.

(c) If F_X has at least one discontinuity, then F_X is not continuous on $\mathbb{R}$ and hence X cannot be a continuous RV. For X to be a discrete RV X, we must have

$$\sum_{x \in D} [F_X(x+) - F_X(x-)] = \sum_{x \in D} \mathbb{P}(X = x) = 1.$$

If the above condition is satisfied, X is a discrete RV. Otherwise, X is not a discrete RV.

Note 1.148. Cantor function (also known as the Devil's Staircase) is an example of a continuous distribution function, which is not absolutely continuous. In this case, the DF F is not representable as $\int_{-\infty}^{x} f(t)\,dt$ for any non-negative integrable function. We do not discuss these types of examples in this course.

Note 1.149. Consider the DF $F : \mathbb{R} \to [0,1]$ considered in Example 1.120 given by

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

As discussed in Example 1.123 and Example 1.134, an RV with DF F is neither discrete nor continuous.

Definition 1.150 (Quantiles and Median for an RV)**.** Let X be an RV with DF F_X. For any $p \in (0,1)$, a number $x \in \mathbb{R}$ is called a quantile of order p if the following inequalities are satisfied, viz.

$$p \leq F_X(x) \leq p + \mathbb{P}(X = x).$$

A quantile of order $\frac{1}{2}$ is called a median.

Note 1.151. A quantile need not be unique. Refer to problem set 4 for explicit examples.

Notation 1.152. We write $\mathfrak{z}_p(X)$ to denote a quantile of order p.

Notation 1.153. The quantiles of order $\frac{1}{4}$ and $\frac{3}{4}$ for an RV X are referred to as the lower and upper quartiles of X, respectively.

Note 1.154. The inequalities mentioned in Definition 1.150 can be restated as

$$\mathbb{P}(X \leq x) \geq p, \quad \mathbb{P}(X \geq x) \geq 1 - p.$$

Note 1.155. Let X be a continuous RV with DF F_X. Then a quantile of order p is a solution to the equation $F_X(x) = p$, since $\mathbb{P}(X = x) = 0, \forall x \in \mathbb{R}$. Moreover, if F_X is strictly increasing, then $\mathfrak{z}_p(X)$ is unique for all $p \in (0,1)$.

Example 1.156. Consider a continuous RV X with DF $F_X : \mathbb{R} \to [0,1]$ given by

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

For any $p \in (0,1)$, the solution to $F_X(x) = p$ is given by $x = p$, i.e. $\mathfrak{z}_p(X) = p$. Moreover, the median is $\mathfrak{z}_{\frac{1}{2}}(X) = \frac{1}{2}$.

We now discuss functions of RVs and their law/distributions.

Remark 1.157 (Function of an RV is an RV). Let $h : \mathbb{R} \to \mathbb{R}$ be a function and let X be an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Since $X : \Omega \to \mathbb{R}$ is a function, we can consider the composition of the functions h and X to obtain another function $h \circ X : \Omega \to \mathbb{R}$ defined by $(h \circ X)(\omega) := h(X(\omega)), \forall \omega \in \Omega$. Since $h \circ X$ is a real valued function defined on Ω with $(\Omega, \mathcal{F}, \mathbb{P})$, $h \circ X$ is an RV defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Notation 1.158. In the setting of the above remark, we shall write $h(X)$ to denote $h \circ X$.

Example 1.159. Let X be an RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and consider the function $h : \mathbb{R} \to \mathbb{R}$ defined by $h(x) = 3x^2 + \sin x + 1, \forall x \in \mathbb{R}$. Then $h(X) = h \circ X$ defined by $(h \circ X)(\omega) := 3X(\omega)^2 + \sin(X(\omega)) + 1, \forall \omega \in \Omega$ is an RV.

Remark 1.160 (DF of a function of an RV). We continue with the notations of Remark 1.157 and are interested in computing the law/distribution of $Y = h(X)$. Using Remark 1.113, we may equivalently, compute the DF of Y and that will identify the required law. Then for any $y \in \mathbb{R}$, we have

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(h(X) \leq y) = \mathbb{P}(h(X) \in (-\infty, y]) = \mathbb{P}(X \in h^{-1}((-\infty, y])),$$

where $h^{-1}((-\infty, y])$ denotes the pre-image of $(-\infty, y]$ under h (see Notation 1.92).

Example 1.161. Let X be a discrete RV with p.m.f.

$$f_X(x) := \begin{cases} \frac{|x|}{110} & \text{if } x \in \{\pm 1, \pm 2, \ldots, \pm 10\} \\ 0, & \text{otherwise} \end{cases}$$

and take $h : \mathbb{R} \to \mathbb{R}$ as $h(x) := |x|, \forall x \in \mathbb{R}$. Note that

$$h^{-1}((-\infty, y]) = \begin{cases} \emptyset, \text{ if } y < 0, \\ \{0\}, \text{ if } y = 0, \\ [-y, y], \text{ if } y > 0. \end{cases}$$

Then the DF of $Y = h(X) = |X|$ is given by

$$\begin{aligned} F_Y(y) &= \mathbb{P}(X \in h^{-1}((-\infty, y])) \\ &= \begin{cases} \mathbb{P}(X \in \emptyset), \text{ if } y < 0, \\ \mathbb{P}(X \in \{0\}), \text{ if } y = 0, \\ \mathbb{P}(X \in [-y, y]), \text{ if } y > 0. \end{cases} \\ &= \begin{cases} 0, \text{ if } y < 0, \\ \mathbb{P}(X = 0), \text{ if } y = 0, \\ \sum_{t \in [-y,y] \cap \{\pm 1, \pm 2, \ldots, \pm 10\}} f_X(t), \text{ if } y > 0. \end{cases} \\ &= \begin{cases} 0, \text{ if } y \leq 0, \\ \sum_{t \in [-y,y] \cap \{\pm 1, \pm 2, \ldots, \pm 10\}} \frac{|t|}{110}, \text{ if } y > 0. \end{cases} \end{aligned}$$

From the structure of the DF we conclude that the RV is discrete. The p.m.f. may be computed using the techniques discussed in earlier lectures.

Example 1.162. Let X be a continuous RV with p.d.f.

$$f_X(x) = \begin{cases} \frac{|x|}{2}, & \text{if } -1 < x < 1 \\ \frac{x}{3}, & \text{if } 1 \le x < 2 \\ 0, & \text{otherwise} \end{cases}$$

and take $h : \mathbb{R} \to \mathbb{R}$ as $h(x) := x^2, \forall x \in \mathbb{R}$. Note that

$$h^{-1}((-\infty, y]) = \begin{cases} \emptyset, \text{ if } y < 0, \\ \{0\}, \text{ if } y = 0, \\ [-\sqrt{y}, \sqrt{y}], \text{ if } y > 0. \end{cases}$$

Then the DF of $Y = h(X) = X^2$ is given by

$$\begin{aligned} F_Y(y) &= \mathbb{P}(X \in h^{-1}((-\infty, y])) \\ &= \begin{cases} \mathbb{P}(X \in \emptyset), \text{ if } y < 0, \\ \mathbb{P}(X \in \{0\}), \text{ if } y = 0, \\ \mathbb{P}(X \in [-\sqrt{y}, \sqrt{y}]), \text{ if } y > 0. \end{cases} \\ &= \begin{cases} 0, \text{ if } y < 0, \\ \mathbb{P}(X = 0), \text{ if } y = 0, \\ \mathbb{P}(\{-\sqrt{y} \le X \le \sqrt{y}\}), \text{ if } y > 0. \end{cases} \\ &= \begin{cases} 0, \text{ if } y < 0, \\ 0, \text{ if } y = 0, \\ \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\, dx, \text{ if } y > 0. \end{cases} \end{aligned}$$

$$= \begin{cases} 0, \text{ if } y < 0, \\ 0, \text{ if } y = 0, \\ \int_{-\sqrt{y}}^{\sqrt{y}} \frac{|x|}{2} dx, \text{ if } 0 \leq y < 1 \\ \int_{-1}^{1} \frac{|x|}{2}\, dx + \int_{1}^{\sqrt{y}} \frac{x}{3} dx, \text{ if } 1 \leq y < 4 \\ 1, \text{ if } y \geq 4 \end{cases}$$

$$= \begin{cases} 0, \text{ if } y \leq 0, \\ \frac{y}{2}, \text{ if } 0 \leq y < 1 \\ \frac{y+2}{6}, \text{ if } 1 \leq y < 4 \\ 1, \text{ if } y \geq 4. \end{cases}$$

From the structure of the DF we conclude that the RV is continuous. The p.d.f. may be computed using the techniques discussed in earlier lectures.

Note 1.163. We continue the discussion in Remark 1.160. In general, we may not be able to reduce/simplify the expression $h^{-1}((-\infty, y])$ further, without additional information about h or X. In what follows, we shall consider the cases where X is discrete or continuous and then attempt to obtain the DF of $h(X)$.

Theorem 1.164. *Let X be a discrete RV with DF F_X, p.m.f. f_X and support S_X. Let $h : \mathbb{R} \to \mathbb{R}$ be a function. Then $Y = h(X)$ is a discrete RV with support $S_Y = h(S_X) := \{h(x) : x \in S_X\}$, p.m.f. f_Y given by*

$$f_Y(y) = \begin{cases} \sum_{x \in h^{-1}(\{y\})} f_X(x), \; \textit{if } y \in S_Y, \\ 0, \; \textit{otherwise} \end{cases}$$

and DF F_Y given by

$$F_Y(y) = \mathbb{P}(Y \leq y) = \sum_{t \in S_Y \cap (-\infty, y]} f_Y(t) = \sum_{\substack{x \in S_X \\ h(x) \leq y}} f_X(x) = \sum_{x \in S_X \cap h^{-1}((-\infty, y])} f_X(x).$$

Proof. Since S_X is a finite or a countably infinite set, the set $h(S_X)$ is also finite or countably infinite. Now,

$$\mathbb{P}(h(X) \in h(S_X)) = \mathbb{P}(X \in h^{-1}(h(S_X))) \geq \mathbb{P}(X \in S_X) = 1$$

and hence $\mathbb{P}(h(X) \in h(S_X)) = 1$. Here, we have used the fact that $h^{-1}(h(S_X)) \supseteq S_X$. Moreover, for any $x \in S_X$,

$$\mathbb{P}(h(X) = h(x)) = \mathbb{P}(X \in h^{-1}(\{h(x)\})) \geq \mathbb{P}(X \in \{x\}) = f_X(x) > 0$$

and hence $Y = h(X)$ is discrete with support $S_Y = h(S_X)$. The expressions for f_Y and F_Y follows from standard arguments. □

Note 1.165. As a consequence of Theorem 1.164, we conclude that the functions of discrete RVs are also discrete RVs.

Note 1.166. In Theorem 1.164, the function h need not be one-to-one or onto and therefore need not have an inverse. This was the same problem encountered in Remark 1.160, which stops us in computing the DF of $h(X)$ for a general RV X.

As a special case of Remark 1.160, we get the next result. We do not give a separate proof, for brevity.

Corollary 1.167. *Continue with the notations of Theorem 1.164. Assume that $h : S_X \to \mathbb{R}$ is one-to-one. Then we have*

$$f_Y(y) = \begin{cases} f_X(h^{-1}(y)), & \text{if } y \in S_Y, \\ 0, & \text{otherwise} \end{cases}$$

where $h^{-1} : h(S_X) \to S_X$ denotes the inverse function of $h : S_X \to \mathbb{R}$.

Example 1.168. Let X be a discrete RV with p.m.f.

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

Consider the RV $Y = X^2$. Here $S_X = \{-2, -1, 0, 1, 2, 3\}$ and $S_Y = \{0, 1, 4, 9\}$. Observe that,

$$\begin{aligned}
\mathbb{P}(Y = 0) &= \mathbb{P}\left(X^2 = 0\right) = \mathbb{P}(X = 0) = \tfrac{1}{7},\\
\mathbb{P}(Y = 1) &= \mathbb{P}\left(X^2 = 1\right) = \mathbb{P}(X \in \{-1, 1\}) = \tfrac{1}{7} + \tfrac{1}{7} = \tfrac{2}{7},\\
\mathbb{P}(Y = 4) &= \mathbb{P}\left(X^2 = 4\right) = \mathbb{P}(X \in \{-2, 2\}) = \tfrac{1}{7} + \tfrac{3}{14} = \tfrac{5}{14}\\
\mathbb{P}(Y = 9) &= \mathbb{P}\left(X^2 = 9\right) = \mathbb{P}(X \in \{-3, 3\}) = 0 + \tfrac{3}{14} = \tfrac{3}{14}.
\end{aligned}$$

Therefore, the p.m.f. of Y is

$$f_Y(y) = \begin{cases} \frac{1}{7}, & \text{if } y = 0\\ \frac{2}{7}, & \text{if } y = 1\\ \frac{5}{14}, & \text{if } y = 4\\ \frac{3}{14}, & \text{if } y = 9\\ 0, & \text{otherwise,} \end{cases}$$

and the DF 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 fact, after identifying S_Y, we could have directly computed the DF F_Y as follows:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \begin{cases} 0, \text{ if } y < 0,\\ \mathbb{P}(Y = 0), \text{ if } 0 \leq y < 1,\\ \mathbb{P}(Y = 0) + \mathbb{P}(Y = 1), \text{ if } 1 \leq y < 4,\\ \mathbb{P}(Y = 0) + \mathbb{P}(Y = 1) + \mathbb{P}(Y = 4), \text{ if } 4 \leq y < 9,\\ 1, \text{ if } y \geq 9. \end{cases}$$

and the p.m.f. f_Y from F_Y using standard techniques discussed in earlier lectures.

Example 1.169. Let X be a discrete RV with p.m.f.

$$f_X(x) = \begin{cases} \frac{x}{55} & \text{if } x \in \{1, 2, \dots, 10\} \\ 0, & \text{otherwise.} \end{cases}$$

Now consider the RV $Y = X^2$. Note that the function $h : \mathbb{R} \to \mathbb{R}$ defined by $h(x) := x^2, \forall x \in \mathbb{R}$ is one-to-one on the support S_X. Here, Y is discrete with support $S_Y = \{1, 4, 9, \cdots, 100\}$ and by Corollary 1.167, the p.m.f. f_Y is given by

$$f_Y(y) = \begin{cases} f_X(\sqrt{y}), \text{ if } y \in S_Y, \\ 0, \text{ otherwise} \end{cases} = \begin{cases} \frac{\sqrt{y}}{55}, \text{ if } y \in S_Y, \\ 0, \text{ otherwise} \end{cases}.$$

The DF F_Y can now be computed from the p.m.f. f_Y using standard techniques.

Now we look at functions of continuous RVs.

Example 1.170. Let X be a continuous RV with p.d.f.

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

and let $Y = [X]$, where $[x]$ denotes the largest integer not exceeding x for $x \in \mathbb{R}$. Note that $S_X = [0, \infty)$. Moreover,

$$\mathbb{P}(Y \in \{0, 1, 2, \dots\}) = \mathbb{P}(X \in S_X) = 1$$

and hence Y is a discrete RV. Now, for $y \in \{0, 1, 2, \dots\}$

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(\{y \leq X < y + 1\}) = \int_y^{y+1} f_X(x)\, dx = \int_y^{y+1} e^{-x}\, dx = \left(1 - e^{-1}\right) e^{-y} > 0.$$

hence Y is a discrete RV with support $S_Y = \{0, 1, 2, \dots\}$ and the above p.m.f. f_Y. Therefore, a function of a continuous RV need not be a continuous RV.

Remark 1.171. Given any continuous RV X and a constant function $h : \mathbb{R} \to \mathbb{R}$ given by $h(x) := c, \forall x \in \mathbb{R}$ for some $c \in \mathbb{R}$, the RV $h(X)$ is discrete. Together with the above example, we may conclude that additional information on h is required before we can conclude that $h(X)$ is continuous.

The next result is stated without proof.

Theorem 1.172. *Let X be a continuous RV with p.d.f. f_X and support S_X. Suppose $\{x \in \mathbb{R} : f_X(x) > 0\} = \cup_{i=1}^k (a_i, b_i)$ and f_X is continuous on each (a_i, b_i). We assume that the intervals (a_i, b_i) are pairwise disjoint.*

Let $h : \mathbb{R} \to \mathbb{R}$ be a function such that on each (a_i, b_i), $h : (a_i, b_i) \to \mathbb{R}$ is strictly monotone and continuously differentiable with inverse function h_i^{-1} for $i = 1, \ldots, k$.

Then $Y = h(X)$ is a continuous RV with support $S_Y = \cup_{i=1}^k [c_i, d_i]$, where $c_i = \min\{h(a_i), h(b_i)\}$ and $d_i = \max\{h(a_i), h(b_i)\}$. The p.d.f. is given by

$$f_Y(y) = \sum_{i=1}^k f_X\left(h_i^{-1}(y)\right) \left|\frac{d}{dy} h_i^{-1}(y)\right| 1_{(c_i,d_i)}(y), y \in \mathbb{R}$$

where $1_{(c_i,d_i)}(y) = 1$ if $y \in (c_i, d_i)$ and 0 otherwise.

Note 1.173. In Theorem 1.172, the function h may be strictly monotone increasing in some (a_i, b_i) and strictly monotone decreasing in other intervals. Moreover, this monotonicity may be verified by looking at the sign of h'. If $h'(x) > 0, \forall x \in (a_i, b_i)$, then h is strictly monotone increasing on (a_i, b_i). If $h'(x) < 0, \forall x \in (a_i, b_i)$, then h is strictly monotone decreasing on (a_i, b_i).

Example 1.174. Let X be a continuous RV with p.d.f.

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

and consider $Y = X^2$. Here, $S_X = [0, \infty)$ and the function $h : \mathbb{R} \to \mathbb{R}$ defined by $h(x) := x^2, \forall x \in \mathbb{R}$ is continuous differentiable on $(0, \infty)$. Moreover, $h'(x) = 2x > 0, \forall x \in (0, \infty)$ and hence h is strictly monotone increasing on $(0, \infty)$. The inverse function is given by $h^{-1}(y) = \sqrt{y}, \forall y \in (0, \infty)$. The p.d.f. f_Y is given by

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

The DF F_Y can now be computed from the p.d.f. f_Y by standard techniques.

Example 1.175. Let X be a continuous RV with p.d.f.

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

and consider $Y = X^2$.

Observe that $\{x \in \mathbb{R} : f_X(x) > 0\} = (-1,0) \cup (0,2)$. Now, $h(x) = x^2$ is strictly decreasing on $(-1,0)$ with inverse function $h_1^{-1}(t) = -\sqrt{t}$; and $h(x) = x^2$ is strictly increasing on $(0,2)$ with inverse function $h_2^{-1}(t) = \sqrt{t}$. Note that $h((-1,0)) = (0,1)$ and $h((0,2)) = (0,4)$. Then, $Y = X^2$ has p.d.f. given by

$$\begin{aligned} f_Y(y) &= f_X(-\sqrt{y}) \left| \frac{d}{dy}(-\sqrt{y}) \right| 1_{(0,1)}(y) + f_X(\sqrt{y}) \left| \frac{d}{dy}(\sqrt{y}) \right| 1_{(0,4)}(y) \\ &= \begin{cases} \frac{1}{2}, & \text{if } 0 < y < 1 \\ \frac{1}{6}, & \text{if } 1 < y < 4 \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

We can compute the DF of Y and verify that this matches with our earlier computation in Example 1.162.

Let X be a discrete (or continuous) RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with DF F_X, support S_X and p.m.f. (or p.d.f.) f_X.

Definition 1.176 (Expectation/Expected value/Mean of the RV X)**.** The Expectation/Expected value/Mean of the RV X, denoted by $\mathbb{E}X$, is defined as the quantity

$$\mathbb{E}[X] := \begin{cases} \sum_{x \in S_X} x f_X(x), & \text{if } \sum_{x \in S_X} |x| f_X(x) < \infty \text{ for discrete } X, \\ \int_{-\infty}^{\infty} x f_X(x)\, dx, & \text{if } \int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty \text{ for continuous } X. \end{cases}$$

Remark 1.177. If the sum or the integral above converges absolutely, we say that the expectation $\mathbb{E}X$ exists or equivalently, $\mathbb{E}X$ is finite. Otherwise, we shall say that the expectation $\mathbb{E}X$ does not exist.

The next result is stated without proof.

Theorem 1.172. *Let X be a continuous RV with p.d.f. f_X and support S_X. Suppose $\{x \in \mathbb{R} : f_X(x) > 0\} = \cup_{i=1}^k (a_i, b_i)$ and f_X is continuous on each (a_i, b_i). We assume that the intervals (a_i, b_i) are pairwise disjoint.*

Let $h : \mathbb{R} \to \mathbb{R}$ be a function such that on each (a_i, b_i), $h : (a_i, b_i) \to \mathbb{R}$ is strictly monotone and continuously differentiable with inverse function h_i^{-1} for $i = 1, \dots, k$.

Then $Y = h(X)$ is a continuous RV with support $S_Y = \cup_{i=1}^k [c_i, d_i]$, where $c_i = \min\{h(a_i), h(b_i)\}$ and $d_i = \max\{h(a_i), h(b_i)\}$. The p.d.f. is given by

$$f_Y(y) = \sum_{i=1}^k f_X\left(h_i^{-1}(y)\right) \left| \frac{d}{dy} h_i^{-1}(y) \right| 1_{(c_i,d_i)}(y), y \in \mathbb{R}$$

where $1_{(c_i,d_i)}(y) = 1$ if $y \in (c_i, d_i)$ and 0 otherwise.

Note 1.173. In Theorem 1.172, the function h may be strictly monotone increasing in some (a_i, b_i) and strictly monotone decreasing in other intervals. Moreover, this monotonicity may be verified by looking at the sign of h'. If $h'(x) > 0, \forall x \in (a_i, b_i)$, then h is strictly monotone increasing on (a_i, b_i). If $h'(x) < 0, \forall x \in (a_i, b_i)$, then h is strictly monotone decreasing on (a_i, b_i).

Example 1.174. Let X be a continuous RV with p.d.f.

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

and consider $Y = X^2$. Here, $S_X = [0, \infty)$ and the function $h : \mathbb{R} \to \mathbb{R}$ defined by $h(x) := x^2, \forall x \in \mathbb{R}$ is continuous differentiable on $(0, \infty)$. Moreover, $h'(x) = 2x > 0, \forall x \in (0, \infty)$ and hence h is strictly monotone increasing on $(0, \infty)$. The inverse function is given by $h^{-1}(y) = \sqrt{y}, \forall y \in (0, \infty)$. The p.d.f. f_Y is given by

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

The DF F_Y can now be computed from the p.d.f. f_Y by standard techniques.

Example 1.175. Let X be a continuous RV with p.d.f.

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

and consider $Y = X^2$.

Observe that $\{x \in \mathbb{R} : f_X(x) > 0\} = (-1, 0) \cup (0, 2)$. Now, $h(x) = x^2$ is strictly decreasing on $(-1, 0)$ with inverse function $h_1^{-1}(t) = -\sqrt{t}$; and $h(x) = x^2$ is strictly increasing on $(0, 2)$ with inverse function $h_2^{-1}(t) = \sqrt{t}$. Note that $h((-1, 0)) = (0, 1)$ and $h((0, 2)) = (0, 4)$. Then, $Y = X^2$ has p.d.f. given by

$$\begin{aligned} f_Y(y) &= f_X(-\sqrt{y}) \left| \frac{d}{dy}(-\sqrt{y}) \right| 1_{(0,1)}(y) + f_X(\sqrt{y}) \left| \frac{d}{dy}(\sqrt{y}) \right| 1_{(0,4)}(y) \\ &= \begin{cases} \frac{1}{2}, & \text{if } 0 < y < 1 \\ \frac{1}{6}, & \text{if } 1 < y < 4 \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

We can compute the DF of Y and verify that this matches with our earlier computation in Example 1.162.

Let X be a discrete (or continuous) RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with DF F_X, support S_X and p.m.f. (or p.d.f.) f_X.

Definition 1.176 (Expectation/Expected value/Mean of the RV X)**.** The Expectation/Expected value/Mean of the RV X, denoted by $\mathbb{E}X$, is defined as the quantity

$$\mathbb{E}[X] := \begin{cases} \sum_{x \in S_X} x f_X(x), & \text{if } \sum_{x \in S_X} |x| f_X(x) < \infty \text{ for discrete } X, \\ \int_{-\infty}^{\infty} x f_X(x)\, dx, & \text{if } \int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty \text{ for continuous } X. \end{cases}$$

Remark 1.177. If the sum or the integral above converges absolutely, we say that the expectation $\mathbb{E}X$ exists or equivalently, $\mathbb{E}X$ is finite. Otherwise, we shall say that the expectation $\mathbb{E}X$ does not exist.

Note 1.178. Note that it is possible to define the expectation $\mathbb{E}X$ through the law/distribution $\mathbb{P}_X$ of X. However, this is beyond the scope of this course.

Example 1.179. Fix $c \in \mathbb{R}$. Let X be a discrete RV with p.m.f.

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} 1, & \text{if } x = c \\ 0, & \text{otherwise.} \end{cases}$$

Such RVs are called constant/degenerate RVs. Here, the support is a singleton set $S_X = \{c\}$ and $\sum_{x \in S_X} |x| f_X(x) = |c| < \infty$ and hence $\mathbb{E}X = \sum_{x \in S_X} x f_X(x) = c$.

Example 1.180. Let X be a discrete RV with p.m.f.

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

Here, the support is $S_X = \{1,2,3,4,5,6\}$, a finite set with all elements positive and hence $\sum_{x \in S_X} |x| f_X(x) = \sum_{x \in S_X} x f_X(x)$ is finite and

$$\mathbb{E}X = \sum_{x \in S_X} x f_X(x) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}.$$

Example 1.181. Let X be a discrete RV with p.m.f.

$$f_X(x) := \begin{cases} \frac{1}{2^x}, \forall x \in \{1,2,3,\cdots\} \\ 0, \text{ otherwise.} \end{cases}$$

Here, the support is $S_X = \{1,2,3,\cdots\}$, the set of natural numbers. To check the existence of $\mathbb{E}X$, we need to check the convergence of the series $\sum_{x \in S_X} |x| f_X(x) = \sum_{x=1}^{\infty} x \frac{1}{2^x}$. Now, the x-th term is $\frac{x}{2^x}$ and

$$\lim_{x \to \infty} \frac{\frac{x+1}{2^{x+1}}}{\frac{x}{2^x}} = \frac{1}{2} < 1.$$

By ratio test, we have the required convergence and the existence of $\mathbb{E}X$ follows.
Observe that

$$\mathbb{E}X = \sum_{x=1}^{\infty} x\frac{1}{2^x} = \frac{1}{2} + \sum_{x=2}^{\infty} x\frac{1}{2^x} = \frac{1}{2} + \sum_{x=1}^{\infty}(x+1)\frac{1}{2^{x+1}} = \frac{1}{2} + \frac{1}{2}\sum_{x=1}^{\infty} x\frac{1}{2^x} + \frac{1}{2} = 1 + \frac{1}{2}\mathbb{E}X,$$

which gives $\mathbb{E}X = 2$.

Note 1.182. It is fact that $\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}$.

Example 1.183. Let X be a discrete RV with p.m.f.

$$f_X(x) := \begin{cases} \frac{3}{\pi^2 x^2}, \forall x \in \{\pm 1, \pm 2, \pm 3, \cdots\} \\ 0, \text{ otherwise.} \end{cases}$$

Here, the support is $S_X = \{\pm 1, \pm 2, \pm 3, \cdots\}$. To check the existence of $\mathbb{E}X$, we need to check the convergence of the series $\sum_{x \in S_X} |x| f_X(x) = 2\sum_{n=1}^{\infty} n\frac{3}{\pi^2 n^2} = \frac{6}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n}$. However, this series diverges and hence $\mathbb{E}X$ does not exist.

Example 1.184. Let X be a continuous RV with the p.d.f.

$$f_X(x) = \begin{cases} 1, \text{ if } 0 < x < 1, \\ 0, \text{ otherwise.} \end{cases}$$

To check the existence of $\mathbb{E}X$, we need to check the existence of $\int_{-\infty}^{\infty} |x| f_X(x)\,dx$. Now,

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx = \int_0^1 x\,dx = \frac{1}{2}$$

and hence $\mathbb{E}X = \frac{1}{2}$.

Example 1.185. Let X be a continuous RV with the p.d.f.

$$f_X(x) = \frac{1}{2}e^{-|x|}, \forall x \in \mathbb{R}.$$

To check the existence of $\mathbb{E}X$, we need to check the existence of $\int_{-\infty}^{\infty} |x| f_X(x)\,dx$. Now,

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx = \int_{-\infty}^{\infty} |x|\frac{1}{2}e^{-|x|}\,dx = \int_0^{\infty} xe^{-x}\,dx = 1 < \infty$$

and hence $\mathbb{E}X$ exists and

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{-\infty}^{\infty} x \frac{1}{2} e^{-|x|}\,dx = 0.$$

Example 1.186. Let X be a continuous RV with the p.d.f.

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \forall x \in \mathbb{R}.$$

To check the existence of $\mathbb{E}X$, we need to check the existence of $\int_{-\infty}^{\infty} |x| f_X(x)\,dx$. Now,

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx = \int_{-\infty}^{\infty} |x| \frac{1}{\pi}\frac{1}{1+x^2}\,dx = \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2}\,dx = \infty$$

and hence $\mathbb{E}X$ does not exist.

Proposition 1.187. *Let X be a discrete or continuous RV such that $\mathbb{E}X$ exists. Then,*

$$\mathbb{E}X = \int_0^{\infty} \mathbb{P}(X > x)\,dx - \int_{-\infty}^{0} \mathbb{P}(X < x)\,dx.$$

Proof. We prove the result when X is continuous. The case for discrete X can be proved in a similar manner. Observe that

$$\begin{aligned}
\mathbb{E}X &= \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{-\infty}^{0} x f_X(x)\,dx + \int_0^{\infty} x f_X(x)\,dx \\
&= -\int_{x=-\infty}^{0}\int_{y=x}^{0} f_X(x)\,dydx + \int_{x=0}^{\infty}\int_{y=0}^{x} f_X(x)\,dydx \\
&= -\int_{y=-\infty}^{0}\int_{x=-\infty}^{y} f_X(x)\,dxdy + \int_{y=0}^{\infty}\int_{x=y}^{\infty} f_X(x)\,dxdy \\
&= \int_0^{\infty} \mathbb{P}(X > y)\,dy - \int_{-\infty}^{0} \mathbb{P}(X < y)\,dy.
\end{aligned}$$

This completes the proof. □

Remark 1.188. (a) Suppose X is discrete or continuous with $\mathbb{P}(X \geq 0) = 1$. Then $\mathbb{P}(X \leq x) = 0, \forall x < 0$ and hence $\mathbb{E}X = \int_0^{\infty} \mathbb{P}(X > x)\,dx$.

(b) Suppose that X is discrete with $\mathbb{P}(X \in \{0, 1, 2, \cdots\}) = 1$. Then $\mathbb{P}(X > x) = \mathbb{P}(X \geq n + 1), \forall x \in [n, n + 1), n \in \{0, 1, 2, \cdots\}$ and hence by part (a),

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > x)\, dx = \sum_{n=0}^{\infty} \mathbb{P}(X \geq n + 1) = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n).$$

Note 1.189 (Expectation of functions of RVs)**.** Given a function $h : \mathbb{R} \to \mathbb{R}$ and an RV X, we have already discussed about the distribution of $Y = h(X)$. If the p.m.f./p.d.f. f_Y is known, we can then consider the existence of $\mathbb{E}Y$ through f_Y, as per Definition 1.176. However, to do this, we first need to compute f_Y from X and then check the relevant existence. In what follows, we discuss the computation of $\mathbb{E}Y = \mathbb{E}h(X)$ directly from X, using the p.m.f./p.d.f. f_X.

Proposition 1.190. *(a) Let X be a discrete RV with p.m.f. f_X and support S_X and let $h : \mathbb{R} \to \mathbb{R}$ be a function. Consider the discrete RV $Y := h(X)$. Then $\mathbb{E}Y$ exists provided $\sum_{x \in S_X} |h(x)| f_X(x) < \infty$ and in this case,*

$$\mathbb{E}Y = \mathbb{E}h(X) = \sum_{x \in S_X} h(x) f_X(x).$$

(b) Let X be a continuous RV with p.d.f. f_X and support S_X and let $h : \mathbb{R} \to \mathbb{R}$ be a function. Consider the RV $Y := h(X)$. Then $\mathbb{E}Y$ exists provided $\int_{-\infty}^{\infty} |h(x)| f_X(x)\, dx < \infty$ and in this case,

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

Proof. We consider the proof for the case when X is discrete. The other case can be proved by similar arguments.

By Theorem 1.164, $Y = h(X)$ is discrete with support $S_Y = h(S_X)$. Now,

$$\sum_{y \in S_Y} |y| f_Y(y) = \sum_{y \in S_Y} |y| \sum_{\{x \in S_X : h(x) = y\}} f_X(x) = \sum_{y \in S_Y} \sum_{\{x \in S_X : h(x) = y\}} |h(x)| f_X(x) = \sum_{x \in S_X} |h(x)| f_X(x).$$

Therefore, $\mathbb{E}Y$ exists provided $\sum_{x \in S_X} |h(x)| f_X(x) < \infty$ and in this case,

$$\mathbb{E}Y = \sum_{y \in S_Y} y f_Y(y) = \sum_{y \in S_Y} y \sum_{\{x \in S_X : h(x) = y\}} f_X(x) = \sum_{x \in S_X} h(x) f_X(x).$$

This completes the proof. □

Note 1.191. If X is discrete with p.m.f. f_X such that $\mathbb{E}X$ exists, then $\mathbb{E}|X| = \sum_{x \in S_X} |x| f_X(x) < \infty$. Similarly, if X is continuous with p.d.f. f_X such that $\mathbb{E}X$ exists, then $\mathbb{E}|X| = \int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty$. Therefore $\mathbb{E}X$ exists if and only if $\mathbb{E}|X| < \infty$. In other words, $\mathbb{E}X$ is finite if and only if $\mathbb{E}|X|$ is finite.

Note 1.192. Fix $a, b \in \mathbb{R}$ with $a \neq 0$. Let X be a discrete/continuous RV with p.m.f./p.d.f. f_X such that $\mathbb{E}X$ exists. Then $Y = aX + b$ is also a discrete/continuous RV. If X is discrete, then

$$\sum_{x \in S_X} |ax + b| f_X(x) \leq |a| \sum_{x \in S_X} |x| f_X(x) + |b| \sum_{x \in S_X} f_X(x) = |a| \mathbb{E}|X| + |b| < \infty$$

and hence $\mathbb{E}(aX + b)$ exists and equals

$$\mathbb{E}(aX + b) = \sum_{x \in S_X} (ax + b) f_X(x) = a \sum_{x \in S_X} x f_X(x) + b \sum_{x \in S_X} f_X(x) = a\, \mathbb{E}X + b.$$

If X is continuous, a similar argument shows $\mathbb{E}(aX + b) = a\, \mathbb{E}X + b$.

Using arguments similar to the above observations, we obtain the next result. We skip the details for brevity.

Proposition 1.193. *Let X be a discrete/continuous RV with p.m.f./p.d.f. f_X.*

(a) Let $h_i : \mathbb{R} \to \mathbb{R}$ be functions and let $a_i \in \mathbb{R}$ for $i = 1, 2, \cdots, n$. Then

$$\mathbb{E}\left(\sum_{i=1}^{n} a_i h_i(X) \right) = \sum_{i=1}^{n} a_i\, \mathbb{E} h_i(X),$$

provided all the expectations above exist.

(b) Let $h_1, h_2 : \mathbb{R} \to \mathbb{R}$ be functions such that $h_1(x) \leq h_2(x), \forall x \in S_X$, where S_X denotes the support of X. Then,

$$\mathbb{E} h_1(X) \leq \mathbb{E} h_2(X),$$

provided all the expectations above exist.

(c) Take $h_1(x) := -|x|, h_2(x) := x, h_3(x) := |x|, \forall x \in \mathbb{R}$. If $\mathbb{E}X$ exists, then

$$-\mathbb{E}|X| \leq \mathbb{E}X \leq \mathbb{E}|X|,$$

i.e. $|\mathbb{E}X| \leq \mathbb{E}|X|$.

(d) If $\mathbb{P}(a \leq X \leq b) = 1$ *for some* $a, b \in \mathbb{R}$*, then* $\mathbb{E}X$ *exists and* $a \leq \mathbb{E}X \leq b$.

Note 1.194. Given an RV X, by choosing different functions $h : \mathbb{R} \to \mathbb{R}$, we obtain several quantities of interest of the form $\mathbb{E}h(X)$.

Definition 1.195 (Moments)**.** The quantity $\mu_r' := \mathbb{E}[X^r]$, if it exists, is called the r-th moment of RV X for $r > 0$.

Definition 1.196 (Absolute Moments)**.** The quantity $\mathbb{E}[|X|^r]$, if it exists, is called the r-th absolute moment of RV X for $r > 0$.

Definition 1.197 (Moments about a point)**.** Let $c \in \mathbb{R}$. The quantity $\mathbb{E}[(X-c)^r]$, if it exists, is called the r-th moment of RV X about c for $r > 0$.

Definition 1.198 (Absolute Moments about a point)**.** Let $c \in \mathbb{R}$. The quantity $\mathbb{E}[|X-c|^r]$, if it exists, is called the r-th absolute moment of RV X about c for $r > 0$.

Note 1.199. It is clear from the definitions above that the usual moments and absolute moments are moments and absolute moments about origin, respectively.

Proposition 1.200. *Let* X *be a discrete/continuous RV such that* $\mathbb{E}|X|^r < \infty$ *for some* $r > 0$*. Then* $\mathbb{E}|X|^s < \infty$ *for all* $0 < s < r$.

Proof. Observe that for all $x \in \mathbb{R}$, we have $|x|^s \leq \max\{|x|^r, 1\} \leq |x|^r + 1$ and hence

$$\mathbb{E}|X|^s \leq \mathbb{E}|X|^r + 1 < \infty.$$

□

Remark 1.201. Suppose that the m-th moment $\mathbb{E}X^m$ of X exists for some positive integer m. Then we have $\mathbb{E}|X|^m < \infty$ (see Note 1.191). By Proposition 1.200, we have $\mathbb{E}|X|^n < \infty$ for all positive integers $n \leq m$ and hence the n-th moment $\mathbb{E}X^n$ exists for X. In particular, the existence of the second moment $\mathbb{E}X^2$ implies the existence of the first moment $\mathbb{E}X$, which is the expectation of X.

Definition 1.202 (Central Moments)**.** Let X be an RV such that $\mu_1' = \mathbb{E}X$ exists. The quantity $\mu_r := \mathbb{E}[(X-\mu_1')^r]$, if it exists, is called the r-th moment of RV X about the mean or r-th central moment of X for $r > 0$.

Definition 1.203 (Variance)**.** The second central moment μ_2 of an RV X, if it exists, is called the variance of X and denoted by $Var(X)$. Note that $Var(X) = \mu_2 = \mathbb{E}\left[(X-\mu_1')^2\right]$.

Remark 1.204. The following are some simple observations about the variance of an RV X.

(a) We have

$$Var(X) = \mathbb{E}\left[(X-\mu_1')^2\right] = \mathbb{E}[X^2+(\mu_1')^2-2\mu_1'X] = \mu_2'-2(\mu_1')^2+(\mu_1')^2 = \mu_2'-(\mu_1')^2 = \mathbb{E}X^2-(\mathbb{E}X)^2.$$

(b) Since the RV $(X-\mu_1')^2$ takes non-negative values, we have $Var(X) = \mathbb{E}(X-\mu_1')^2 \geq 0$.

(c) We have $(\mathbb{E}X)^2 \leq \mathbb{E}X^2$.

(d) $Var(X) = 0$ if and only if $\mathbb{P}(X = \mu_1') = 1$. (see problem set 5).

(e) For any $a, b \in \mathbb{R}$, we have $Var(aX+b) = a^2 Var(X)$.

(f) Let $Var(X) > 0$. Then $Y := \frac{X-\mathbb{E}X}{\sqrt{Var(X)}}$ has the property that $\mathbb{E}Y = 0$ and $Var(Y) = 1$.

Definition 1.205 (Standard Deviation)**.** The quantity $\sigma(X) = \sqrt{Var(X)}$ is defined to be the standard deviation of X.

Example 1.206. Let X be a discrete RV with p.m.f.

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

Here, existence of $\mu_1' = \mathbb{E}X$ and $\mu_2' = \mathbb{E}X^2$ can be established by standard calculations. Moreover,

$$\mathbb{E}X = \sum_{x \in S_X} x f_X(x) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$$

and

$$\mathbb{E}X^2 = \sum_{x \in S_X} x^2 f_X(x) = \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) = \frac{91}{6}.$$

Variance can now be computed using the relation $Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.

Example 1.207. In Example 1.184, we had shown $\mathbb{E}X = \frac{1}{2}$, where X is a continuous RV with the p.d.f.

$$f_X(x) = \begin{cases} 1, \text{ if } 0 < x < 1, \\ 0, \text{ otherwise.} \end{cases}$$

Now, $\mathbb{E}X^2 = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^1 x^2\,dx = \frac{1}{3}$. Then $Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$.

Note 1.208. We are familiar with the Laplace transform of a given real-valued function defined on $\mathbb{R}$. We also know that under certain conditions, the Laplace transform of a function determines the function almost uniquely. In probability theory, the Laplace transform of a p.m.f./p.d.f. of a random variable X plays an important role.

Let X be a discrete/continuous RV defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with DF F_X, p.m.f./p.d.f. f_X and support S_X.

Definition 1.209 (Moment Generating Function (MGF))**.** We say that the moment generating function (MGF) of X exists, denoted by M_X and equals $M_X(t) := \mathbb{E}e^{tX}$, provided $\mathbb{E}e^{tX}$ exists for all $t \in (-h, h)$, for some $h > 0$.

Note 1.210. Observe that $e^x > 0, \forall x \in \mathbb{R}$.

Note 1.211. If X is discrete/continuous with p.m.f./p.d.f. f_X, then following the definition of an expectation of an RV, we write

$$M_X(t) = \mathbb{E}e^{tX} = \begin{cases} \sum\limits_{x \in S_X} e^{tx} f_X(x), & \text{if } \sum\limits_{x \in S_X} e^{tx} f_X(x) < \infty \text{ for discrete } X, \forall t \in (-h, h) \text{ for some } h > 0 \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx, & \text{if } \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx < \infty \text{ for continuous } X, t \in (-h, h) \text{ for some } h > 0. \end{cases}$$

In this case, we shall say that the MGF M_X exists on $(-h, h)$.

Remark 1.212. (a) $M_X(0) = 1$ and hence $A := \left\{t \in \mathbb{R} : \mathbb{E}[e^{tX}] \text{ is finite}\right\} \neq \emptyset$.

(b) $M_X(t) > 0\ \forall t \in A$, with A as above.

(c) For $c \in \mathbb{R}$, consider the constant/degenerate RV X given by the p.m.f. (see Example 1.179)

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

Here, the support is $S_X = \{c\}$ and $M_X(t) = \mathbb{E}e^{tX} = \sum_{x \in S_X} e^{tx} f_X(x) = e^{tc}$ exists for all $t \in \mathbb{R}$.

(d) Suppose the MGF M_X exists on $(-h, h)$. Take constants $c, d \in \mathbb{R}$ with $c \neq 0$. Then, the RV $Y = cX + d$ is discrete/continuous, according to X being discrete/continuous and moreover,

$$M_Y(t) = \mathbb{E}e^{t(cX+d)} = e^{td} M_X(ct)$$

exists for all $t \in (-\frac{h}{|c|}, \frac{h}{|c|})$.

Note 1.213. The MGF can be used to compute the moments of an RV and this is the motivation behind the term 'Moment Generating Function'. This result is stated below. We skip the proof for brevity.

Theorem 1.214. *Let X be an RV with MGF M_X which exists on $(-h, h)$ for some $h > 0$. Then, we have the following results.*

(a) $\mu_r' = \mathbb{E}[X^r]$ is finite for each $r \in \{1, 2, \ldots\}$.

(b) $\mu_r' = \mathbb{E}[X^r] = M_X^{(r)}(0)$, where $M_X^{(r)}(0) = \left[\frac{d^r}{dt^r} M_X(t)\right]_{t=0}$ is the r-th derivative of $M_X(t)$ at the point 0 for each $r \in \{1, 2, \ldots\}$.

(c) M_X has the following Maclaurin's series expansion around $t = 0$ of the following form $M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_r'$ with $t \in (-h, h)$.

Proposition 1.215. *Continue with the notations and assumptions of Theorem 1.214 and define $\psi_X : (-h, h) \to \mathbb{R}$ by $\psi_X(t) := \ln M_X(t), t \in (-h, h)$. Then*

$$\mu_1' = \mathbb{E}[X] = \psi_X^{(1)}(0) \quad \text{and} \quad \mu_2 = Var(X) = \psi_X^{(2)}(0),$$

where $\psi_X^{(r)}$ denotes the r-th $(r \in \{1, 2\})$ derivative of ψ_X.

Proof. We have, for $t \in (-h, h)$

$$\psi_X^{(1)}(t) = \frac{M_X^{(1)}(t)}{M_X(t)} \quad \text{and} \quad \psi_X^{(2)}(t) = \frac{M_X(t)M_X^{(2)}(t) - \left(M_X^{(1)}(t)\right)^2}{\left(M_X(t)\right)^2}.$$

Evaluating the above equalities at $t = 0$ give the required results. □

Example 1.216. Let X be a discrete RV with p.m.f.

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

where $\lambda > 0$. We have

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda\left(e^t - 1\right)} \ \forall\ t \in \mathbb{R}$$

since $A = \left\{t \in \mathbb{R} : \mathbb{E}\left(e^{tX}\right) < \infty\right\} = \mathbb{R}$. Now,

$$M_X^{(1)}(t) = \lambda e^t e^{\lambda\left(e^t-1\right)} \quad \text{and} \quad M_X^{(2)}(t) = \lambda e^t e^{\lambda\left(e^t-1\right)}\left(1 + \lambda e^t\right) \ \forall\ t \in \mathbb{R}.$$

Then,

$$\mu_1' = \mathbb{E}(X) = M_X^{(1)}(0) = \lambda,\ \mu_2' = \mathbb{E}(X^2) = M_X^{(2)}(0) = \lambda(1+\lambda),\ Var(X) = \mu_2 = \mu_2' - (\mu_1')^2 = \lambda.$$

Again, for $t \in \mathbb{R}$, $\psi_X(t) = \ln\left(M_X(t)\right) = \lambda\left(e^t - 1\right)$, which yields $\psi_X^{(1)}(t) = \psi_X^{(2)}(t) = \lambda e^t, \forall t \in \mathbb{R}$. Then, $\mu_1' = \mathbb{E}(X) = \lambda, \mu_2 = Var(X) = \lambda$. Higher order moments can be calculated by looking at higher order derivatives of M_X.

Example 1.217. Let X be a continuous RV with p.d.f.

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

We have

$$M_X(t) = \mathbb{E}\left(e^{tX}\right) = \int_0^{\infty} e^{tx}e^{-x}\,\mathrm{d}x = \int_0^{\infty} e^{-(1-t)x}\,\mathrm{d}x = (1-t)^{-1} < \infty, \text{ if } t < 1.$$

In particular, M_X exists on $(-1,1)$ and $A = \left\{t \in \mathbb{R} : \mathbb{E}\left(e^{tX}\right) < \infty\right\} = (-\infty, 1) \supset (-1,1)$. Now,

$$M_X^{(1)}(t) = (1-t)^{-2} \quad \text{and} \quad M_X^{(2)}(t) = 2(1-t)^{-3}, t < 1.$$

Then,

$$\mu_1' = \mathbb{E}(X) = M_X^{(1)}(0) = 1,\ \mu_2' = \mathbb{E}(X^2) = M_X^{(2)}(0) = 2,\ Var(X) = \mu_2 = \mu_2' - (\mu_1')^2 = 1.$$

Again, for $t < 1$, $\psi_X(t) = \ln\left(M_X(t)\right) = -\ln(1-t)$, which yields $\psi_X^{(1)}(t) = \frac{1}{1-t}, \psi_X^{(2)}(t) = \frac{1}{(1-t)^2}, \forall t < 1$. Then, $\mu_1' = \mathbb{E}(X) = 1, \mu_2 = Var(X) = 1$.

Now, consider the Maclaurin's series expansion for M_X around $t = 0$. We have

$$M_X(t) = (1-t)^{-1} = \sum_{r=0}^{\infty} t^r, \forall t \in (-1,1)$$

and hence $\mu_r' = r!$, which is the coefficient of $\frac{t^r}{r!}$ in the above power series.

Example 1.218. Let X be a continuous RV with p.d.f.

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, \forall x \in \mathbb{R}.$$

As observed earlier in Example 1.186, $\mathbb{E}X$ does not exist. Since the existence of moments is a necessary condition for the existence of MGF, we conclude that the MGF does not exist for this RV X.

Remark 1.219 (Identically distributed RVs). Let X and Y be two RVs, possibly defined on different probability spaces.

(a) Recall from Remark 1.112 that their law/distribution may be the same and in this case, we have $F_X = F_Y$, i.e. $F_X(x) = F_Y(x), \forall x \in \mathbb{R}$. The statement '$X$ and Y are equal in law/distribution' is equivalent to 'X and Y are identically distributed'.

(b) Recall from Remark 1.113 that the DF uniquely identifies the law/distribution, i.e. if $F_X = F_Y$, then X and Y are identically distributed.

(c) Suppose X and Y are discrete RVs. Recall from Remark 1.127, the p.m.f. is uniquely determined by the DF and vice versa. In the case of discrete RVs, X and Y are identically distributed if and only if the p.m.f.s are equal (i.e., $f_X = f_Y$).

(d) Suppose X and Y are continuous RVs. Recall from Note 1.138 that the p.d.f.s in this case are uniquely identified upto sets of 'length 0'. We may refer to such an almost equal p.d.f. as a 'version of a p.d.f.'. Recall from Note 1.141, the p.d.f. is uniquely determined by the DF and vice versa. In the case of continuous RVs, X and Y are identically distributed if and only if the p.d.f.s are versions of each other. In other words, X and Y are identically distributed if and only if there exist versions f_X and f_Y of the p.d.f.s such that $f_X = f_Y$, i.e. $f_X(x) = f_Y(x), \forall x \in \mathbb{R}$.

(e) Suppose X and Y are identically distributed and let $h : \mathbb{R} \to \mathbb{R}$ be a function. Then we have that the RVs $h(X)$ and $h(Y)$ are identically distributed. In particular, $\mathbb{E}h(X) = \mathbb{E}h(Y)$, provided one of the expectations exists.

(f) Suppose X and Y are identically distributed. By (e), X^2 and Y^2 are identically distributed and $\mathbb{E}X^2 = \mathbb{E}Y^2$, provided one of the expectations exists. More generally, the n-th moments $\mathbb{E}X^n$ and $\mathbb{E}Y^n$ of X and Y are the same, provided they exist.

(g) There are examples where $\mathbb{E}X^n = \mathbb{E}Y^n, \forall n = 1, 2, \cdots$, but X and Y are not identically distributed. We may discuss such an example later in this course. Consequently, the moments do not uniquely identify the distribution. Under certain sufficient conditions on the moments, such as the Carleman's condition, it is however possible to uniquely identify the distribution. This is beyond the scope of this course.

(h) Suppose X and Y are identically distributed and suppose that the MGF M_X exists on $(-h, h)$ for some $h > 0$. By the above observation (e), the MGF M_Y exists and $M_X = M_Y$, i.e. $M_X(t) = M_Y(t), \forall t \in (-h, h)$.

(i) We now state a result without proof. Suppose the MGFs M_X and M_Y exist. If $M_X(t) = M_Y(t), \forall t \in (-h, h)$, then X and Y are identically distributed. Therefore, the MGF uniquely identifies the distribution.

Notation 1.220. We write $X \stackrel{d}{=} Y$ to denote that X and Y are identically distributed.

Example 1.221. If Y is an RV with the MGF $M_Y(t) = (1-t)^{-1}, \forall t \in (-1,1)$, then by Example 1.217, we conclude that Y is a continuous RV with p.d.f.

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

Example 1.222. If X is a discrete RV with support S_X and p.m.f. f_X, then the MGF M_X is of the form

$$M_X(t) = \sum_{x \in S_X} e^{tx} f_X(x).$$

We can also make a converse statement. Since the MGF uniquely identifies a distribution, if an MGF is given by a sum of the above form, we can immediately identify the corresponding discrete RV with its support and p.m.f.. For example, if $M_X(t) = \frac{1}{2} + \frac{1}{3}e^t + \frac{1}{6}e^{-t}$, then X is discrete with the p.m.f.

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

Notation 1.223. We may refer to expectations of the form $\mathbb{E}e^{tX}$ as exponential moments of the RV X.

Definition 1.224 (Symmetric Distribution)**.** An RV X is said to have a symmetric distribution about a point $\mu \in \mathbb{R}$ if $X - \mu \overset{d}{=} \mu - X$.

Proposition 1.225. *Let X be an RV which is symmetric about 0.*

(a) If X is discrete, then the p.m.f. f_X has the property that $f_X(x) = f_X(-x), \forall x \in \mathbb{R}$. Further, $\mathbb{E}X^n = 0, \forall n = 1, 3, 5, \cdots$, provided the moments exist.

(b) If X is continuous, then the p.d.f. f_X has the property that $f_X(x) = f_X(-x), \forall x \in \mathbb{R}$. Further, $\mathbb{E}X^n = 0, \forall n = 1, 3, 5, \cdots$, provided the moments exist.

Proof. We prove the statement when X is a continuous RV. The proof for the case when X is discrete is similar.

If X is symmetric about 0, then $X \overset{d}{=} -X$ and hence for any $x \in \mathbb{R}$, $F_X(x) = F_{-X}(x)$ and hence $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(-X \leq x) = \mathbb{P}(X \geq -x) = 1 - F_X(-x)$. This implies $f_X(x) = f_X(-x), \forall x \in \mathbb{R}$.

Assume that the moments in question exist. Then $\mathbb{E}X^n = \int_{-\infty}^{\infty} x^n f_X(x)\,dx = 0$, since the function $x \mapsto x^n f_X(x)$ is odd. □

Remark 1.226. Let X be a continuous RV, which is symmetric about $\mu \in \mathbb{R}$. As argued in the above proposition, we have for all $x \in \mathbb{R}$

$$F_X(\mu + x) = \mathbb{P}(X \leq \mu + x) = \mathbb{P}(X - \mu \leq x) = \mathbb{P}(\mu - X \leq x) = \mathbb{P}(X - \mu \geq -x) = 1 - F_X(\mu - x)$$

and hence $f_X(\mu+x) = f_X(\mu-x), \forall x$. Conversely, given a continuous RV X such that $f_X(\mu+x) = f_X(\mu - x), \forall x$ for some $\mu \in \mathbb{R}$, we have $F_{X-\mu} = F_{\mu-X}$ and hence X is symmetric about μ.

We now look at some special examples of discrete RVs.

Example 1.227 (Degenerate RV)**.** We have already mentioned this example earlier in Example 1.179. Fix $c \in \mathbb{R}$. Say that X is degenerate at c if its distribution is given by the p.m.f.

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} 1, & \text{if } x = c \\ 0, & \text{otherwise.} \end{cases}$$

This is a discrete RV with support $S_X = \{c\}$. As computed earlier, $\mathbb{E}X = c$. We also have $\mathbb{E}X^n = c^n, \forall n \geq 1$ and $M_X(t) = e^{tc}, \forall t \in \mathbb{R}$. Note that $Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 0$.

Remark 1.228 (Bernoulli Trial). Suppose that a random experiment has exactly two outcomes, identified as a ‘success’ and a ‘failure’. For example, while tossing a coin, we may think of obtaining a head as a success and a tail as a failure. Here, the sample space is $\Omega = \{Success, Failure\}$. A single trial of such an experiment is referred to as a Bernoulli trial. In this case, $Probability(\{Success\}) = 1 - Probability(\{Failure\})$. If we define an RV $X : \Omega \to \mathbb{R}$ by $X(Success) = 1$ and $X(Failure) =$

0, then X is a discrete RV with p.m.f.

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} 1 - Probability(\{Success\}), \text{ if } x = 0, \\ Probability(\{Success\}), \text{ if } x = 1, \\ 0, \text{ otherwise.} \end{cases}$$

If $Probability(\{Success\}) = 0$ or $Probability(\{Failure\}) = 0$, then X is degenerate at 0 or 1, respectively. The case when $Probability(\{Success\}) \in (0, 1)$ is therefore of interest.

Example 1.229 (Bernoulli(p) RV)**.** Let $p \in (0, 1)$. An RV X is said to follow Bernoulli(p) distribution or equivalently, X is a Bernoulli(p) RV if its distribution is given by the p.m.f.

$$f_X(x) = \begin{cases} 1 - p, \text{ if } x = 0, \\ p, \text{ if } x = 1, \\ 0, \text{ otherwise.} \end{cases}$$

In relation with the Bernoulli trial described above, p may be treated as the probability of success. Here, $\mathbb{E}X = p$, $\mathbb{E}X^2 = p$, $Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = p - p^2 = p(1-p)$, $M_X(t) = 1 - p + pe^t, t \in \mathbb{R}$. By standard arguments, we can establish the existence of these moments.

Notation 1.230. We may write $X \sim Bernoulli(p)$ to mean that X is a Bernoulli(p) RV. Similar notations shall be used for other RVs and their distributions.

Example 1.231 (Binomial(n, p) RV)**.** Fix a positive integer n and let $p \in (0, 1)$. By the Binomial theorem, we have

$$1 = [p + (1-p)]^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

and hence the function $f : \mathbb{R} \to [0, 1]$ given by

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, \text{ if } x \in \{0, 1, \cdots, n\}, \\ 0, \text{ otherwise.} \end{cases}$$

is a p.m.f.. An RV X is said to follow Binomial(n,p) distribution or equivalently, X is a Binomial(n,p) RV if its distribution is given by the above p.m.f.. Here,

$$\mathbb{E}X = \sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} = np\,[p+(1-p)]^{n-1} = np,$$

and

$$\mathbb{E}X(X-1) = \sum_{k=0}^{n} k(k-1)\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = n(n-1)p^2\sum_{k=2}^{n}\binom{n-2}{k-2}p^{k-2}(1-p)^{n-k} = n(n-1)p^2.$$

Then $\mathbb{E}X^2 = \mathbb{E}X(X-1) + \mathbb{E}X = n(n-1)p^2 + np$ and $Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p)$. Also

$$M_X(t) = \mathbb{E}e^{tX} = \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} = (1-p+pe^t)^n, \forall t \in \mathbb{R}.$$

By standard arguments, we can establish the existence of these moments.

Note 1.232. Observe that Binomial$(1,p)$ distribution is the same as Bernoulli(p) distribution. We shall explore the connection between Binomial and Bernoulli distributions later in the course.

Remark 1.233 (Factorial moments). In the computation for $\mathbb{E}X^2$ for $X \sim Binomial(n,p)$, we first computed $\mathbb{E}X(X-1)$, which is easy to compute. It turns out that expectations of the form $\mathbb{E}X(X-1)$, $\mathbb{E}X(X-1)(X-2)$ etc. are often easy to compute for integer valued RVs X. We refer to such expectations as factorial moments of X.

Remark 1.234 (Symmetry of Binomial$(n,\frac{1}{2})$ distribution). Let $X \sim Binomial(n,p)$ and let $Y := n - X$. Since $M_X(t) = (1-p+pe^t)^n, \forall t \in \mathbb{R}$, we have

$$M_Y(t) = \mathbb{E}e^{tY} = \mathbb{E}e^{t(n-X)} = e^{-nt}M_X(-t) = e^{nt}(1-p+pe^{-t})^n = (p+(1-p)e^t)^n.$$

Since MGFs determine the distribution, we conclude that $Y \sim Binomial(n, 1-p)$. In particular, if $p = \frac{1}{2}$, then $Y = n - X \stackrel{d}{=} X \sim Binomial(n,\frac{1}{2})$. Rewriting the relation, we get $\frac{n}{2} - X \stackrel{d}{=} X - \frac{n}{2}$. Therefore, $X \sim Binomial(n,\frac{1}{2})$ is symmetric about $\frac{n}{2}$.

We now look at more examples of discrete RVs. Later in the course, we shall discuss their motivation through various random experiments.

Example 1.235 (Uniform RVs with support on a finite set)**.** Consider a discrete RV X with support $S_X = \{x_1, x_2, \cdots, x_n\}$ and p.m.f. $f_X : \mathbb{R} \to [0,1]$ given by

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

We had considered the case $S_X = \{1, 2, \cdots, 6\}$ in Example 1.180 and computed the expectation. In the general setting, we have

$$\mathbb{E}X = \frac{1}{n} \sum_{x \in S_X} x, \quad \mathbb{E}X^2 = \frac{1}{n} \sum_{x \in S_X} x^2, \quad M_X(t) = \mathbb{E}e^{tX} = \frac{1}{n} \sum_{x \in S_X} e^{tx}, \forall t \in \mathbb{R}$$

and hence $Var(X)$ can be computed by the formula $\mathbb{E}X^2 - (\mathbb{E}X)^2$. By standard arguments, we can establish the existence of these moments.

Example 1.236 (Poisson (λ) RV)**.** Fix $\lambda > 0$. Note that $e^\lambda = \sum_{k=0}^\infty \frac{\lambda^k}{k!}$ and hence the function $f : \mathbb{R} \to [0,1]$ given by

$$f(x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!}, & \text{if } x \in \{0, 1, 2, \cdots\}, \\ 0, & \text{otherwise.} \end{cases}$$

is a p.m.f.. An RV X is said to follow Poisson(λ) distribution or equivalently, X is a Poisson(λ) RV if its distribution is given by the above p.m.f.. Recall that we have already computed the following $\mathbb{E}X = \lambda, Var(X) = \lambda$ and $M_X(t) = e^{\lambda\left(e^t - 1\right)}, \forall t \in \mathbb{R}$ in Example 1.216. As done for the case of Binomial(n, p) RVs, we can compute factorial moments. For example,

$$\mathbb{E}X(X-1) = \sum_{k=0}^\infty k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} = \lambda^2 \sum_{k=2}^\infty e^{-\lambda} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.$$

In fact, $\mathbb{E}X(X-1)\cdots(X-(n-1)) = \lambda^n$ for all $n \geq 1$.

Example 1.237 (Geometric (p) RV)**.** Fix $p \in (0,1)$. Note that $\sum_{k=0}^\infty p(1-p)^k = 1$ and hence the function $f : \mathbb{R} \to [0,1]$ given by

$$f(x) = \begin{cases} p(1-p)^x, & \text{if } x \in \{0, 1, 2, \cdots\}, \\ 0, & \text{otherwise.} \end{cases}$$

is a p.m.f.. An RV X is said to follow Geometric(p) distribution or equivalently, X is a Geometric(p) RV if its distribution is given by the above p.m.f.. Let us compute the MGF. Here,

$$M_X(t) = \mathbb{E}e^{tX} = \sum_{k=0}^{\infty} e^{tk} p(1-p)^k = \frac{p}{1-(1-p)e^t},$$

for all t such that $0 < (1-p)e^t < 1$ or equivalently, $t < \ln\left(\frac{1}{1-p}\right)$. Looking at the derivatives of M_X and evaluating at $t = 0$, we have $\mathbb{E}X = \frac{1-p}{p}$ and $Var(X) = \frac{1-p}{p^2}$.

We now look at special examples of continuous RVs.

Example 1.238 (Uniform(a,b) RV)**.** Fix $a, b \in \mathbb{R}$ with $a < b$. An RV X is said to follow Uniform(a,b) distribution or equivalently, X is a Uniform(a,b) RV if its distribution is given by the p.d.f.

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

We had considered the case $a = 0, b = 1$ in Example 1.184 and computed the expectation. In the general setting, we have

$$\mathbb{E}X = \int_a^b \frac{x}{b-a}\,dx = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}, \quad \mathbb{E}X^2 = \int_a^b \frac{x^2}{b-a}\,dx = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}$$

and hence $Var(X)$ can be computed by the formula $\mathbb{E}X^2 - (\mathbb{E}X)^2$. The MGF is given by

$$\mathbb{E}e^{tX} = \int_a^b \frac{e^{tx}}{b-a}\,dx = \begin{cases} \frac{e^{tb}-e^{ta}}{t(b-a)}, & \text{if } t \neq 0, \\ 1, & \text{if } t = 0. \end{cases}$$

By standard arguments, we can establish the existence of these moments. Further, observe that $f_X(\frac{a+b}{2} - x) = f_X(\frac{a+b}{2} + x), \forall x \in \mathbb{R}$. Using Remark 1.226, we conclude that X is symmetric about its mean.

Example 1.239 (Cauchy(μ,θ) RV)**.** Let $\theta > 0$ and $\mu \in \mathbb{R}$. An RV X is said to follow Cauchy(μ,θ) distribution if its distribution is given by the p.d.f.

$$f_X(x) = \frac{\theta}{\pi}\frac{1}{\theta^2 + (x-\mu)^2}, \forall x \in \mathbb{R}.$$

The fact that f_X is a p.d.f. is easy to check. Set $y = \frac{x-\mu}{\theta}$ and observe that

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+y^2}\,dy = \frac{2}{\pi}\int_{0}^{\infty} \frac{1}{1+y^2}\,dy = \frac{2}{\pi}\tan^{-1}(y) \mid_0^{\infty} = 1.$$

We have already considered the case $\mu = 0, \theta = 1$ in Example 1.186 and Example 1.218, where we have seen that $\mathbb{E}X$ and the MGF do not exist for this distribution. In the general setting, note that $\frac{X-\mu}{\theta} \sim Cauchy(0,1)$ and by a similar argument, we can show that $\mathbb{E}X$ and MGF do not exist. Moreover, $f_X(\mu + x) = f_X(\mu - x), \forall x \in \mathbb{R}$ and using Remark 1.226, we conclude that X is symmetric about μ.

Example 1.240 (Exponential(λ) RV)**.** Let $\lambda > 0$. Note that $\int_0^\infty \exp(-\frac{x}{\lambda})\,dx = \lambda$ and hence the function $f : \mathbb{R} \to [0, \infty)$ given by

$$f(x) = \begin{cases} \frac{1}{\lambda}\exp(-\frac{x}{\lambda}), & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

is a p.d.f.. An RV X is said to follow Exponential(λ) distribution or equivalently, X is an Exponential(λ) RV if its distribution is given by the above p.d.f.. We have already considered the case $\lambda = 1$ in Example 1.217, where we computed the moments and the MGF. Following similar arguments, in the general setting we have

$$\mathbb{E}X^n = \lambda^n n!, \quad Var(X) = \lambda^2, \quad M_X(t) = (1-\lambda t)^{-1}, \forall t < \frac{1}{\lambda}.$$

By standard arguments, we can establish the existence of these moments.

Definition 1.241 (Gamma function)**.** Recall that the integral $\int_0^\infty x^{\alpha-1}e^{-x}\,dx$ exists if and only if $\alpha > 0$. On $(0,\infty)$, consider the function $\alpha \mapsto \int_0^\infty x^{\alpha-1}e^{-x}\,dx$. It is called the Gamma function and the value at any $\alpha > 0$ is denoted by $\Gamma(\alpha)$.

Remark 1.242. We recall some important properties of the Gamma function.

(a) For $\alpha > 0$, we have $\Gamma(\alpha) > 0$.

(b) $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, if $\alpha > 1$.

(c) $\Gamma(1) = \int_0^\infty e^{-x}\,dx = 1$ and hence using (b), $\Gamma(n) = (n-1)!$ for all positive integers n.

(d) $\Gamma(\frac{1}{2}) = \int_0^\infty \frac{1}{\sqrt{x}} e^{-x}\, dx = \sqrt{\pi}$. Putting $x = \frac{y^2}{2}$, this relation may be rewritten as

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{2} \int_0^\infty \exp\left(-\frac{y^2}{2}\right)\, dy = \sqrt{\pi}.$$

(e) Fix $\beta > 0$. Putting $x = \frac{y}{\beta}$, in the integral for $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$, we get $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} \beta^{-\alpha} \exp(-\frac{y}{\beta})\, dy$.

Example 1.243 (Gamma(α, β) RV)**.** Fix $\alpha > 0, \beta > 0$. By the properties of the Gamma function described above, the function $f : \mathbb{R} \to [0, \infty)$ defined by

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

is a p.d.f.. An RV X is said to follow Gamma(α, β) distribution or equivalently, X is a Gamma(α, β) RV if its distribution is given by the above p.d.f.. Note that for $\alpha = 1$, we get back the p.d.f. for an Exponential(β) RV (see Example 1.240), i.e. Gamma$(1, \beta)$ distribution is the same as Exponential(β) distribution. For general $\alpha > 0, \beta > 0$, we have

$$\mathbb{E}X = \alpha\beta, \quad Var(X) = \alpha\beta^2, \quad M_X(t) = (1 - \beta t)^{-\alpha}, \forall t < \frac{1}{\beta}.$$

By standard arguments, we can establish the existence of these moments.

Example 1.244 (Normal(μ, σ^2) RV)**.** Fix $\mu \in \mathbb{R}, \sigma > 0$. Note that $\Gamma(\frac{1}{2}) = \int_0^\infty \frac{1}{\sqrt{t}} e^{-t}\, dt = \sqrt{\pi}$ (see Remark 1.242). Putting $t = \frac{y^2}{2}$ and after suitable manipulation, we have $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \exp\left(-\frac{y^2}{2}\right)\, dy = 1$. Putting $y = \frac{1}{\sigma}(x - \mu)$ (equivalently, $x = \sigma y + \mu$), we have

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^\infty \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\, dx = 1.$$

Therefore, the function $f : \mathbb{R} \to [0, \infty)$ defined by

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

is a p.d.f.. An RV X is said to follow Normal(μ, σ^2) distribution or equivalently, X is a Normal(μ, σ^2) RV, denoted by $X \sim N(\mu, \sigma^2)$ if its distribution is given by the above p.d.f.. If $X \sim N(\mu, \sigma^2)$, from our above discussion we conclude that $Y = \frac{X-\mu}{\sigma} \sim N(0,1)$. Now,

$$\begin{aligned}
M_Y(t) = \mathbb{E}e^{tY} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ty} \exp\left(-\frac{y^2}{2}\right) dy \\
&= \exp\left(\frac{t^2}{2}\right) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(y-t)^2}{2}\right) dy \\
&= \exp\left(\frac{t^2}{2}\right), \forall t \in \mathbb{R}.
\end{aligned}$$

In particular, $\psi_Y(t) = \ln M_Y(t) = \frac{t^2}{2}, \forall t \in \mathbb{R}$ with $\psi'(t) = t, \psi''(t) = 1, \forall t \in \mathbb{R}$. Evaluating at $t = 0$, by Proposition 1.215 we conclude that $\mathbb{E}Y = 0$ and $Var(Y) = 1$. But $X = \sigma Y + \mu$ and hence $\mathbb{E}X = \mu, Var(X) = \sigma^2$. This yields the interpretation of the parameters μ and σ in the distribution of X. Further, $M_X(t) = \mathbb{E}e^{tX} = \mathbb{E}e^{t(\sigma Y + \mu)} = e^{\mu t} M_Y(\sigma t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2), \forall t \in \mathbb{R}$.

Definition 1.245 (Standard Normal RV)**.** We say X is a Standard Normal RV if $X \sim N(0,1)$, i.e. $\mathbb{E}X = 0$ and $Var(X) = 1$.

Notation 1.246. Normal RVs are also referred to as Gaussian RVs and Normal distribution as Gaussian distribution.

Remark 1.247 (Symmetry of Gaussian Distribution). If $X \sim N(\mu, \sigma^2)$, note that $f_X(\mu + x) = f_X(\mu - x), \forall x \in \mathbb{R}$ and using Remark 1.226, we conclude that X is symmetric about its mean μ.

Remark 1.248 (Moments of a Standard Normal RV). Let $X \sim N(0,1)$. Then X is symmetric about 0 and using Proposition 1.225, we conclude $\mathbb{E}X^n = 0$ for all odd positive integers n. If n is an even positive integer, then $n = 2m$ for some positive integer m and

$$\begin{aligned}
\mathbb{E}X^n &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^n \exp\left(-\frac{x^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2m} \exp\left(-\frac{x^2}{2}\right) dx
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^\infty x^{2m} \exp\left(-\frac{x^2}{2}\right) dx \\
&= \frac{2^m}{\sqrt{\pi}} \int_0^\infty y^{m-\frac{1}{2}} \exp\left(-y\right) dy, \text{ (putting } y = \frac{x^2}{2}) \\
&= \frac{2^m}{\sqrt{\pi}} \Gamma(m + \frac{1}{2}) \\
&= 2^m \left(m - \frac{1}{2}\right) \times \cdots \times \frac{3}{2} \times \frac{1}{2} \\
&= (2m-1) \times \cdots \times 3 \times 1 =: (2m-1)!!,
\end{aligned}
$$

where we have used the properties of the Gamma function. In particular, $\mathbb{E}X^4 = 3$.

Definition 1.249 (Beta function)**.** Recall that the integral $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$ exists if and only if $\alpha > 0$ and $\beta > 0$. On $(0,\infty) \times (0,\infty)$, consider the function $(\alpha,\beta) \mapsto \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$. It is called the Beta function and the value at any (α,β) is denoted by $B(\alpha,\beta)$.

Remark 1.250. Note that for $\alpha > 0, \beta > 0$, we have $B(\alpha,\beta) > 0$ and $B(\alpha,\beta) = B(\beta,\alpha)$. Moreover,

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

Example 1.251 (Beta(α,β) RV)**.** Fix $\alpha > 0, \beta > 0$. By the properties of the Beta function described above, the function $f : \mathbb{R} \to [0,\infty)$ defined by

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

is a p.d.f.. An RV X is said to follow Beta(α,β) distribution or equivalently, X is a Beta(α,β) RV if its distribution is given by the above p.d.f.. If $\alpha = \beta$, then $f(1-x) = f(x), \forall x \in \mathbb{R}$ and hence $X \overset{d}{=} 1 - X$. Then, $X - \frac{1}{2} \overset{d}{=} \frac{1}{2} - X$, i.e., X is symmetric about $\frac{1}{2}$. For all $\alpha, \beta, r > 0$, we have

$$\mathbb{E}X^r = \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha+r-1}(1-x)^{\beta-1}\, dx = \frac{B(\alpha+r,\beta)}{B(\alpha,\beta)}$$

and in particular,

$$\mathbb{E}X = \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{\alpha}{\alpha+\beta}$$

and

$$\mathbb{E}X^2 = \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}.$$

Then

$$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

We now study important inequalities in connection with moments of RVs and probabilities of events involving the RVs. Given any RV X, we shall always assume that it is either discrete with p.m.f. f_X or continuous with p.d.f. f_X, if not stated otherwise.

Note 1.252. At times, it is possible to compute the moments of an RV, but the computation of probability of certain events involving the RV may be difficult. The inequalities, that we are going to study, give us estimates of the probabilities in question.

Theorem 1.253. *Let X be an RV such that X is non-negative (i.e. $\mathbb{P}(X \geq 0) = 1$). Suppose that $\mathbb{E}X$ exists. Then for any $c > 0$, we have*

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

Proof. We discuss the proof when X is a continuous RV with p.d.f. f_X. The case when X is discrete can be proved using similar arguments.

For $x < 0$, we have $F_X(x) = \mathbb{P}(X \leq x) \leq \mathbb{P}(X < 0) = 1 - \mathbb{P}(X \geq 0) = 0$ and hence $f_X(x) = 0, \forall x < 0$. Then,

$$\mathbb{E}X = \int_0^\infty x f_X(x)\,dx \geq \int_c^\infty x f_X(x)\,dx \geq c\int_c^\infty f_X(x)\,dx = c\,\mathbb{P}(X \geq c).$$

This completes the proof. □

Note 1.254. Under the assumptions of Theorem 1.253, we have $\mathbb{E}X \geq 0$.

The following special cases of Theorem 1.253 are quite useful in practice.

Corollary 1.255. *(a) Let X be an RV and let $h : \mathbb{R} \to [0, \infty)$ be a function such that $\mathbb{E}h(X)$ exists. Then for any $c > 0$, we have*

$$\mathbb{P}(h(X) \geq c) \leq \frac{\mathbb{E}h(X)}{c}.$$

(b) Let X be an RV and let $h : \mathbb{R} \to [0, \infty)$ be a strictly increasing function such that $\mathbb{E}h(X)$ exists. Then for any $c > 0$, we have

$$\mathbb{P}(X \geq c) = \mathbb{P}(h(X) \geq h(c)) \leq \frac{\mathbb{E}h(X)}{h(c)}.$$

(c) Let X be an RV such that $\mathbb{E}X$ exists, i.e. $\mathbb{E}|X| < \infty$. Considering the RV $|X|$, for any $c > 0$ we have

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}|X|}{c}.$$

(d) (Markov's inequality) Let $r > 0$ and let X be an RV such that $\mathbb{E}|X|^r < \infty$. Then for any $c > 0$, we have

$$\mathbb{P}(|X| \geq c) = \mathbb{P}(|X|^r \geq c^r) \leq c^{-r}\, \mathbb{E}|X|^r.$$

(e) (Chernoff's inequality) Let X be an RV with $\mathbb{E}e^{\lambda X} < \infty$ for some $\lambda > 0$. Then for any $c > 0$, we have

$$\mathbb{P}\{X \geq c\} = \mathbb{P}\{e^{\lambda X} \geq e^{\lambda c}\} \leq e^{-\lambda c}\, \mathbb{E}e^{\lambda X}.$$

Note 1.256. Let X be an RV with finite second moment, i.e. $\mu_2' = \mathbb{E}X^2 < \infty$. By Remark 1.201, the first moment $\mu_1' = \mathbb{E}X$ exists. Hence

$$\mathbb{E}(X - c)^2 = \mathbb{E}[X^2 + c^2 - 2cX] = \mathbb{E}X^2 + c^2 - 2c\,\mathbb{E}X = \mu_2' + c^2 - 2c\,\mu_1' < \infty$$

Therefore, all second moments of X about any point $c \in \mathbb{R}$ exists. In particular, $Var(X) = \mathbb{E}(X - \mu_1')^2 < \infty$. By a similar argument, for any RV X with finite variance, we have $\mathbb{E}X^2 < \infty$.

The next result is a special case of Markov's inequality.

Corollary 1.257 (Chebyshev's inequality)**.** *Let X be an RV with finite second moment (equivalently, finite variance). Then*

$$\mathbb{P}[|X-\mu_1'| \geq c] \leq \frac{1}{c^2}\mathbb{E}(X-\mu_1')^2 = \frac{1}{c^2}Var(X).$$

Remark 1.258. Another form of the above result is also useful. Under the same assumptions, for any $\epsilon > 0$ we have

$$\mathbb{P}[|X-\mu_1'| \geq \epsilon\,\sigma(X)] \leq \frac{1}{\epsilon^2},$$

where $\sigma(X)$ is the standard deviation of X. This measures the spread/deviation of the distribution (of X) about the mean in multiples of the standard deviation. The smaller the variance, lesser the spread.

Remark 1.259. In general, bounds in Theorem 1.253 or in Markov/Chebyshev's inequalities are very conservative. However, they can not be improved further. To see this, consider a discrete RV X with p.m.f. given by

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

Then $\mathbb{P}(X \geq 1) = \frac{1}{4} = \mathbb{E}X$, which is sharp. If we consider

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

then, $\mathbb{P}(X \geq 1) = \frac{1}{4} < \frac{1}{2} = \mathbb{E}X$.

Definition 1.260 (Convex functions)**.** Let I be an open interval in $\mathbb{R}$. We say that a function $h : I \to \mathbb{R}$ is convex on I if

$$h(\alpha x + (1-\alpha)y) \leq \alpha h(x) + (1-\alpha)h(y), \forall \alpha \in (0,1), \forall x, y \in I.$$

We say that h is strictly convex on I if the above inequality is strict for all x, y and α.

We state the following result from Real Analysis without proof.

Theorem 1.261. *Let I be an open interval in $\mathbb{R}$ and let $h : I \to \mathbb{R}$ be a function.*

(a) If h is convex on $\mathbb{R}$, then h is continuous on $\mathbb{R}$.

(b) Let h be twice differentiable on I. Then,

(i) h is convex if and only if $h''(x) \geq 0, \forall x \in I$.

(ii) h is strictly convex if and only if $h''(x) > 0, \forall x \in I$.

The following result is stated without proof.

Theorem 1.262 (Jensen's Inequality)**.** *Let I be an interval in $\mathbb{R}$ and let $h : I \to \mathbb{R}$ be a convex function. Let X be an RV with support $S_X \subseteq I$. Then,*

$$h(\mathbb{E}X) \leq \mathbb{E}h(X),$$

provided the expectations exist. If h is strictly convex, then the inequality above is strict unless X is a degenerate RV.

Remark 1.263. Some special cases of Jensen's inequality are of interest.

(a) Consider $h(x) = x^2, \forall x \in \mathbb{R}$. Here, $h''(x) = 2 > 0, \forall x$ and hence h is convex on $\mathbb{R}$. Then $(\mathbb{E}X)^2 \leq \mathbb{E}X^2$, provided the expectations exist. We had seen this inequality earlier in Remark 1.204.

(b) For any integer $n \geq 2$, consider the function $h(x) = x^n$ on $[0, \infty)$. Here, $h''(x) = n(n-1)x^{n-2} \geq 0, \forall x \in (0, \infty)$ and hence h is convex. Then $(\mathbb{E}|X|)^n \leq \mathbb{E}|X|^n$, provided the expectations exist.

(c) Consider $h(x) = e^x, \forall x \in \mathbb{R}$. Here, $h''(x) = e^x > 0, \forall x$ and hence h is convex on $\mathbb{R}$. Then $e^{\mathbb{E}X} \leq \mathbb{E}e^X$, provided the expectations exist.

(d) Consider any RV X with $\mathbb{P}(X > 0) = 1$ and look at $h(x) := -\ln x, \forall x \in (0, \infty)$. Then $h''(x) = \frac{1}{x^2} > 0, \forall x \in (0, \infty)$ and hence h is convex. Then $-\ln(\mathbb{E}X) \leq \mathbb{E}(-\ln X)$, i.e. $\ln(\mathbb{E}X) \geq \mathbb{E}(\ln X)$, provided the expectations exist.

(e) Consider any RV X with $\mathbb{P}(X > 0) = 1$. Then $\mathbb{P}(\frac{1}{X} > 0) = 1$ and hence by (d), $-\ln(\mathbb{E}\frac{1}{X}) \leq \mathbb{E}(-\ln \frac{1}{X}) = \mathbb{E}(\ln X)$. Then $(\mathbb{E}\frac{1}{X})^{-1} = e^{-\ln(\mathbb{E}\frac{1}{X})} \leq e^{\mathbb{E}(\ln X)} \leq \mathbb{E}X$, by (c). This inequality holds, provided all the expectations exist. We may think of $\mathbb{E}X$ as the arithmetic mean (A.M.) of X, $e^{\mathbb{E}(\ln X)}$ as the geometric mean (G.M.) of X, and $\frac{1}{\mathbb{E}[\frac{1}{X}]}$ as the harmonic mean (H.M.) of X. The inequality obtained here is related to the classical A.M.-G.M.-H.M. inequality (see problem set 6).

Note 1.264 (Why should we look at multiple RVs together?)**.** Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ associated with a random experiment $\mathcal{E}$. As motivated earlier, an RV associates some numerical quantity to each of the outcomes of the experiment. Such numerical quantities help us in the understanding of characteristics of the outcomes. However, it is important to note that, in practice, we may be interested in looking at these characteristics of the outcomes at the same time. This also allows us to see if the characteristics in question may be related. If we perform the random experiment separately for each of these characteristics, then there is also the issue of cost and time associated with the repeated performance of the experiment. Keeping this in mind, we now choose to consider multiple characteristics of the outcomes at the same time. This leads us to the concept of Random Vectors, which allows us to look at multiple RVs at the same time.

Example 1.265. Consider the random experiment of rolling a standard six-sided die three times. Here, the sample space is

$$\Omega = \{(i, j, k) : i, j, k \in \{1, 2, 3, 4, 5, 6\}\}.$$

Suppose we are interested in the sum of the first two rolls and the sum of all rolls. These characteristics of the outcomes can be captured by the RVs $X, Y : \Omega \to \mathbb{R}$ defined by $X((i, j, k)) := i + j$ and $Y((i, j, k)) := i + j + k$ for all $(i, j, k) \in \Omega$. If we look at X and Y simultaneously, we may comment on whether a 'large' value for X implies a 'large' Y and vice versa.

Definition 1.266 (Random Vector)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A function $X = (X_1, X_2, \cdots, X_p) : \Omega \to \mathbb{R}^p$ is called a p-dimensional random vector (or simply, a random vector, if the dimension p is clear from the context). Here, the component functions are denoted by $X_1, X_2, \cdots, X_p$ and each of these are real valued functions defined on the sample space Ω and hence are RVs.

Note 1.267. A 1-dimensional random vector, by definition, is exactly an RV. A p-dimensional random vector is made up of p components, each of which are RVs. Keeping this connection in mind, we repeat the steps of our analysis as done for RVs.

Notation 1.268 (Pre-image of a set under an $\mathbb{R}^p$-valued function)**.** Let Ω be a non-empty set and let $X : \Omega \to \mathbb{R}^p$ be a function. Given any subset A of $\mathbb{R}^p$, we consider the subset $X^{-1}(A)$ of Ω defined by

$$X^{-1}(A) := \{\omega \in \Omega : X(\omega) \in A\}.$$

The set $X^{-1}(A)$ shall be referred to as the pre-image of A under the function X. We shall suppress the symbols ω and use the following notation for convenience, viz.

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} = (X \in A).$$

Notation 1.269. As discussed for RVs, we now consider the following set function in relation to a given p-dimensional random vector. Given a random vector defined on $(\Omega, \mathcal{F}, \mathbb{P})$, consider the set function $\mathbb{P}_X(A) := \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A)$ for all subsets A of $\mathbb{R}^p$. We shall write $\mathbb{B}_p$ to denote the power set of $\mathbb{R}^p$.

Following arguments similar to Proposition 1.103, we get the next result. The proof is skipped for brevity.

Proposition 1.270. *Let X be a p-dimensional random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, the set function $\mathbb{P}_X$ is a probability function/measure defined on the collection $\mathbb{B}_p$, i.e. $(\mathbb{R}^p, \mathbb{B}_p, \mathbb{P}_X)$ is a probability space.*

Definition 1.271 (Induced Probability Space and Induced Probability Measure)**.** If X is a p-dimensional random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the probability function/measure $\mathbb{P}_X$ on $\mathbb{B}_p$ is referred to as the induced probability function/measure induced by X. In this case, $(\mathbb{R}^p, \mathbb{B}_p, \mathbb{P}_X)$ is referred to as the induced probability space induced by X.

Notation 1.272. We shall call $\mathbb{P}_X$ as the joint law or joint distribution of the random vector X.

We have found that the DF of an RV identifies the law/distribution of the RV. Motivated by this fact, we now consider a similar function for random vectors.

Definition 1.273 (Joint Distribution function (Joint DF) and Marginal Distribution function (Marginal DF))**.** Let $X = (X_1, X_2, \cdots, X_p) : \Omega \to \mathbb{R}^p$ be a p-dimensional random vector.

(a) The joint DF of X is a function $F_X : \mathbb{R}^p \to [0, 1]$ defined by

$$\begin{aligned} F_X(x_1, x_2, \cdots, x_p) &:= \mathbb{P}_X((-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_p]) \\ &= \mathbb{P}(X \in \prod_{j=1}^{p}(-\infty, x_j]) \\ &= \mathbb{P}((X_1, X_2, \cdots, X_p) \in \prod_{j=1}^{p}(-\infty, x_j]) \\ &= \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_p \leq x_p), \forall x = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^p. \end{aligned}$$

(b) The joint DF of any subset of the RVs $X_1, X_2, \cdots, X_p$ is called a marginal DF of the random vector X.

Note 1.274. Let $X = (X_1, X_2, X_3) : \Omega \to \mathbb{R}^3$ be a 3-dimensional random vector. Then the DF F_{X_2} of X_2 and the joint DF F_{X_1,X_3} of X_1 & X_3 are marginal DFs of the random vector X.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Unless stated otherwise, RVs and random vectors shall be defined on this probability space.

Note 1.275. Recall that for an RV Y, we have $F_Y(b) - F_Y(a) = \mathbb{P}(a < Y \leq b) \geq 0$ for all $a, b \in \mathbb{R}$ with $a < b$.

Proposition 1.276. *Let $X = (X_1, X_2) : \Omega \to \mathbb{R}^2$ be a 2-dimensional random vector. Let $a_1 < b_1, a_2 < b_2$. Then,*

$$\begin{aligned} F_X(b_1, b_2) - F_X(a_1, b_2) - F_X(b_1, a_2) + F_X(a_1, a_2) &= \mathbb{P}(X \in (a_1, b_1] \times (a_2, b_2]) \\ &= \mathbb{P}(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2) \\ &\geq 0. \end{aligned}$$

Proof. Consider the events $A_1 := (X_1 \leq a_1, X_2 \leq b_2)$ and $A_2 := (X_1 \leq b_1, X_2 \leq a_2)$. Note that

$$(X_1 \leq b_1, X_2 \leq b_2) \cap (a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2)^c = A_1 \cup A_2.$$

Now, $(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2) \subseteq (X_1 \leq b_1, X_2 \leq b_2)$ and hence

$$\begin{aligned}
&\mathbb{P}((X_1 \leq b_1, X_2 \leq b_2) \cap (a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2)^c) \\
&= \mathbb{P}(X_1 \leq b_1, X_2 \leq b_2) - \mathbb{P}(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2) \\
&= F_X(b_1, b_2) - \mathbb{P}(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2).
\end{aligned}$$

By the inclusion-exclusion principle (see Proposition 1.61)

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) = F_X(a_1, b_2) + F_X(b_1, a_2) - F_X(a_1, a_2).$$

The result follows. □

For higher dimensions, the above result has an appropriate extension. To state this, we first need some notations.

Notation 1.277. Let $\prod_{j=1}^{p}(a_j, b_j]$ be a rectangle in $\mathbb{R}^p$. Observe that the co-ordinates of the vertices are made up of either a_j or b_j for each $j = 1, 2, \cdots, p$. Let Δ_k^p denote the set of vertices where exactly k many a_j's appear. Then the complete set of vertices is $\cup_{k=0}^{p} \Delta_k^p$. For example,

$$\Delta_0^2 = \{(b_1, b_2)\}, \quad \Delta_1^2 = \{(a_1, b_2), (b_1, a_2)\}, \quad \Delta_2^2 = \{(a_1, a_2)\}.$$

Proposition 1.276 can now be generalized to higher dimensions as follows. We skip the details of the proof for brevity.

Imp

Proposition 1.278. *Let $X = (X_1, X_2, \cdots, X_p) : \Omega \to \mathbb{R}^p$ be a p-dimensional random vector. Let $a_1 < b_1, a_2 < b_2, \cdots, a_p < b_p$. Then,*

$$\mathbb{P}(X \in \prod_{j=1}^{p}(a_j, b_j]) = \sum_{k=0}^{p}(-1)^k \sum_{x \in \Delta_k^p} F_X(x) = \mathbb{P}(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2, \ldots, a_p < X_p \leq b_p) \geq 0.$$

Proposition 1.279 (Computation of Marginal DFs from Joint DF)**.** *Let $X = (X_1, X_2, \cdots, X_p) : \Omega \to \mathbb{R}^p$ be a p-dimensional random vector. Fix $1 \leq j \leq p$. Then, for all $x \in \mathbb{R}$ we have*

$$\begin{aligned} F_{X_j}(x) &= \lim_{\substack{t_k \to \infty \\ k \in \{1, \cdots, j-1, j+1, \cdots, p\}}} F_X(t_1, \cdots, t_{j-1}, x, t_{j+1}, \cdots, t_p) \\ &= \lim_{t \to \infty} F_X(\underbrace{t, \cdots, t}_{j-1 \textit{ times}}, x, \underbrace{t, \cdots, t}_{p-j \textit{ times}}) \\ &=: F_X(\underbrace{\infty, \cdots, \infty}_{j-1 \textit{ times}}, x, \underbrace{\infty, \cdots, \infty}_{p-j \textit{ times}}). \end{aligned}$$

Proof. As in the proof of Theorem 1.115, using Proposition 1.110, we have

$$\begin{aligned} &\lim_{\substack{t_k \to \infty \\ k \in \{1, \cdots, j-1, j+1, \cdots, p\}}} F_X(t_1, \cdots, t_{j-1}, x, t_{j+1}, \cdots, t_p) \\ &= \lim_{\substack{t_k \to \infty \\ k \in \{1, \cdots, j-1, j+1, \cdots, p\}}} \mathbb{P}_X((-\infty, t_1] \times \cdots (-\infty, t_{j-1}] \times (-\infty, x] \times (-\infty, t_{j-1}] \times \cdots \times (-\infty, t_p]) \\ &= \lim_{n \to \infty} \mathbb{P}_X((-\infty, n] \times \cdots (-\infty, n] \times (-\infty, x] \times (-\infty, n] \times \cdots \times (-\infty, n]) \\ &= \mathbb{P}_X(\mathbb{R} \times \cdots \times \mathbb{R} \times (-\infty, x] \times \mathbb{R} \times \cdots \times \mathbb{R}) \\ &= \mathbb{P}(X_1 \in \mathbb{R}, \cdots, X_{j-1} \in \mathbb{R}, X_j \in (-\infty, x], X_{j+1} \in \mathbb{R}, \cdots, X_p \in \mathbb{R}) \\ &= \mathbb{P}(X_j \in (-\infty, x]) = F_{X_j}(x). \end{aligned}$$

This completes the proof. □

Remark 1.280. Using Proposition 1.279, we can compute the DFs of each component RVs from the joint DF of a random vector. More generally, the higher dimensional marginal DFs can be computed from the joint DF in a similar manner. For example, if $X = (X_1, X_2, \cdots, X_p)$ is a p-dimensional random vector, then

$$F_{X_1, X_2}(x_1, x_2) = \lim_{t \to \infty} F_X(x_1, x_2, \underbrace{t, \cdots, t}_{p-2 \text{ times}}) =: F_X(x_1, x_2, \underbrace{\infty, \cdots, \infty}_{p-2 \text{ times}}).$$

The joint DF of a random vector has properties similar to the DF of an RV. Compare the next result with Theorem 1.115.

Theorem 1.281. *Let $X = (X_1, X_2, \cdots, X_p) : \Omega \to \mathbb{R}^p$ be a p-dimensional random vector with joint DF F_X. Then,*

(a) F_X is non-decreasing in the sense of Proposition 1.278, i.e. for $a_1 < b_1, a_2 < b_2, \cdots, a_p < b_p$ we have

$$\sum_{k=0}^{p} (-1)^k \sum_{x \in \Delta_k^p} F_X(x) \geq 0.$$

(b) F_X is jointly right continuous in the co-ordinates, i.e.

$$\lim_{\substack{h_k \downarrow 0 \\ k \in \{1,2,\cdots,p\}}} F_X(x_1 + h_1, x_2 + h_2, \cdots, x_p + h_p) = F_X(x_1, x_2, \cdots, x_p).$$

In particular, F_X is right continuous in each co-ordinate, keeping other co-ordinates fixed.

(c) We have

$$\lim_{\substack{x_k \to \infty \\ k \in \{1,2,\cdots,p\}}} F_X(x_1, x_2, \cdots, x_p) = 1.$$

(d) For any fixed $j \in \{1, 2, \cdots, p\}$ and $x_1, x_2, \cdots, x_{j-1}, x_{j+1}, \cdots, x_p \in \mathbb{R}$, we have

$$\lim_{x_j \to -\infty} F_X(x_1, x_2, \cdots, x_p) = 0.$$

Proof. Statement (a) is already mentioned in Proposition 1.278.

Proofs of (b), (c) and (d) follow from Proposition 1.110, similar to the proof of Theorem 1.115. We only prove (b) to illustrate the idea.

$$\begin{aligned}
&\lim_{\substack{h_k \downarrow 0 \\ k \in \{1,2,\cdots,p\}}} F_X(x_1 + h_1, x_2 + h_2, \cdots, x_p + h_p) \\
&= \lim_{\substack{h_k \downarrow 0 \\ k \in \{1,2,\cdots,p\}}} \mathbb{P}_X((-\infty, x_1 + h_1] \times (-\infty, x_2 + h_2] \times \cdots \times (-\infty, x_p + h_p]) \\
&= \lim_{n \to \infty} \mathbb{P}_X(\left(-\infty, x_1 + \frac{1}{n}\right] \times \left(-\infty, x_2 + \frac{1}{n}\right] \times \cdots \times \left(-\infty, x_p + \frac{1}{n}\right]) \\
&= \mathbb{P}_X((-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_p])
\end{aligned}$$

$$= F_X(x_1, x_2, \cdots, x_p).$$

□

The next theorem, an analogue of Theorem 1.116, is stated without proof. The arguments required to prove this statement is beyond the scope of this course.

Theorem 1.282. *Any function $F : \mathbb{R}^p \to [0,1]$ satisfying the properties in Theorem 1.281 is the joint DF of some p-dimensional random vector.*

Note 1.283. Using arguments similar to above discussion, it is immediate that the joint DF of a random vector is non-decreasing in each co-ordinate, keeping other co-ordinates fixed.

Definition 1.284 (Mutually Independent RVs). Let $\mathcal{I}$ be a non-empty indexing set (can be finite, countably infinite or uncountable). We say that a collection of RVs $\{X_\alpha : \alpha \in \mathcal{I}\}$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is mutually independent (or simply, independent) if for all finite subcollections $\{X_{\alpha_1}, X_{\alpha_2}, \cdots, X_{\alpha_n}\}$ we have

Imp

$$F_{X_{\alpha_1}, X_{\alpha_2}, \cdots, X_{\alpha_n}}(x_1, x_2, \cdots, x_n) = \prod_{j=1}^{n} F_{X_{\alpha_j}}(x_j), \forall x_1, x_2, \cdots, x_n \in \mathbb{R}.$$

Notation 1.285. If a collection of RVs $\{X_\alpha : \alpha \in \mathcal{I}\}$ is independent, we may also say that the RVs $X_\alpha, \alpha \in \mathcal{I}$ are independent.

Proposition 1.286. *The RVs $X_1, X_2, \cdots, X_p$, with $p \geq 2$, are independent if and only if*

$$F_{X_1, X_2, \cdots, X_p}(x_1, x_2, \cdots, x_p) = \prod_{j=1}^{p} F_{X_j}(x_j), \forall x_1, x_2, \cdots, x_p \in \mathbb{R}.$$

Proof. If the RVs $X_1, X_2, \cdots, X_p$ are independent, then the relation involving the joint DF follows from the definition.

Conversely, let $\mathcal{J} \subset \{1, 2, \cdots, p\}$. We would like to show that the subcollection $\{X_j : j \in \mathcal{J}\}$ is independent. Let Y be the $|\mathcal{J}|$-dimensional random vector with the component RVs $X_j, j \in \mathcal{J}$. Then F_Y is a joint DF of Y as well as a marginal DF of X. Then by Remark 1.280, for all $y \in \mathbb{R}^{|\mathcal{J}|}$,

$$F_Y(y) = \lim_{\substack{x_j \to \infty, j \notin \mathcal{J} \\ x_j = y_j, j \in \mathcal{J}}} F_X(x) = \lim_{\substack{x_j \to \infty, j \notin \mathcal{J} \\ x_j = y_j, j \in \mathcal{J}}} \prod_{j \notin \mathcal{J}} F_{X_j}(x_j) \prod_{j \in \mathcal{J}} F_{X_j}(x_j) = \prod_{j \in \mathcal{J}} F_{X_j}(y_j).$$

This shows that the subcollection $\{X_j : j \in \mathcal{J}\}$ is independent and the proof is complete. □

Remark 1.287. It follows from the definition that if a collection of RVs $\{X_\alpha : \alpha \in \mathcal{I}\}$ is independent, then any subcollection of RVs $\{X_\alpha : \alpha \in \mathcal{J}\}$, with $\mathcal{J} \subset \mathcal{I}$ is also independent.

Definition 1.288 (Pairwise Independent RVs)**.** Let $\mathcal{I}$ be a non-empty indexing set (can be finite, countably infinite or uncountable). We say that a collection of RVs $\{X_\alpha : \alpha \in \mathcal{I}\}$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is pairwise independent if for all distinct indices $\alpha, \beta \in \mathcal{I}$, the subcollection $\{X_\alpha, X_\beta\}$ is independent, i.e.

$$F_{X_\alpha, X_\beta}(x_1, x_2) = F_{X_\alpha}(x_1) F_{X_\beta}(x_2), \forall x_1, x_2 \in \mathbb{R}.$$

Note 1.289. So far, we have not discussed examples of random vectors. In fact, as considered for RVs, we shall consider special classes of random vectors and explicit examples shall then be discussed.

Definition 1.290 (Discrete Random Vector)**.** A random vector $X = (X_1, X_2, \cdots, X_p)$ is said to be a discrete random vector if there exists a finite or countably infinite set $S \subset \mathbb{R}^p$ such that

$$1 = \mathbb{P}_X(S) = \mathbb{P}(X \in S) = \sum_{x \in S} \mathbb{P}_X(\{x\}) = \sum_{x \in S} \mathbb{P}(X = x)$$

and $\mathbb{P}(X = x) > 0, \forall x \in S$. In this situation, we refer to the set S as the support of the discrete random vector X.

Definition 1.291 (Joint Probability Mass Function for a discrete random vector)**.** Let $X = (X_1, X_2, \cdots, X_p)$ be a discrete random vector with support S_X. Consider the function $f_X : \mathbb{R}^p \to \mathbb{R}$ defined by

$$f_X(x) := \begin{cases} \mathbb{P}(X = x), \text{ if } x \in S_X, \\ 0, \text{ if } x \in S_X^c. \end{cases}$$

This function f_X is called the joint probability mass function (joint p.m.f.) of the random vector X.

Remark 1.292. Let $X = (X_1, X_2, \cdots, X_p)$ be a discrete random vector with joint DF F_X, joint p.m.f. f_X and support S_X. Then, similar to the p.m.f. for RVs, we have the following observations.

(a) The joint p.m.f. $f_X : \mathbb{R}^p \to \mathbb{R}$ is a function such that

$$f_X(x) = 0, \forall x \in S_X^c, \quad f_X(x) > 0, \forall x \in S_X, \quad \sum_{x \in S_X} f_X(x) = 1.$$

(b) $\mathbb{P}_X(S_X^c) = 1 - \mathbb{P}_X(S_X) = 0$. In particular, $\mathbb{P}(X = x) = f_X(x) = 0, \forall x \in S_X^c$.

(c) Since $\mathbb{P}_X(S_X) = 1$, for any $A \subseteq \mathbb{R}^p$ we have,

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}_X(A \cap S_X) = \sum_{x \in A \cap S_X} \mathbb{P}(X = x) = \sum_{x \in A \cap S_X} f_X(x).$$

Since S_X is finite or countably infinite, the set $A \cap S_X$ is also finite or countably infinite.

(d) By (c), for any $x = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^p$, we consider $A = \prod_{j=1}^p (-\infty, x_j]$, we obtain

$$\begin{aligned} F_X(x) &= \mathbb{P}_X\left(\prod_{j=1}^p (-\infty, x_j]\right) \\ &= \mathbb{P}(X_1 \le x_1, X_2 \le x_2, \cdots, X_p \le x_p) \\ &= \sum_{y \in S_X \cap \prod_{j=1}^p (-\infty, x_j]} f_X(y). \end{aligned}$$

Therefore, the joint p.m.f. f_X is uniquely determined by the joint DF F_X and vice versa.

(e) To study a discrete random vector X, we may study any one of the following three quantities, viz. the joint law/distribution $\mathbb{P}_X$, the joint DF F_X or the joint p.m.f. f_X.

(f) For any $j \in \{1, 2, \cdots, p\}$, for $x_j \in \mathbb{R}$

$$\begin{aligned} F_{X_j}(x_j) &= \mathbb{P}(X_j \in (-\infty, x_j]) \\ &= \mathbb{P}(X_1 \in \mathbb{R}, \cdots, X_{j-1} \in \mathbb{R}, X_j \in (-\infty, x_j], X_{j+1} \in \mathbb{R}, \cdots, X_p \in \mathbb{R}) \\ &= \mathbb{P}_X(\mathbb{R} \times \cdots \times \mathbb{R} \times (-\infty, x_j] \times \mathbb{R} \times \cdots \times \mathbb{R}) \\ &= \sum_{y \in S_X \cap \mathbb{R} \times \cdots \times \mathbb{R} \times (-\infty, x_j] \times \mathbb{R} \times \cdots \times \mathbb{R}} f_X(y) \end{aligned}$$

$$= \sum_{\substack{y \in S_X \\ y_j \leq x_j}} f_X(y).$$

Consider $g_j : \mathbb{R} \to \mathbb{R}$ defined by $g_j(x) := \sum\limits_{\substack{y \in S_X \\ y_j = x}} f_X(y)$. It is immediate that g_j satisfies the properties of a p.m.f. and $F_{X_j}(x_j) = \sum_{z \leq x_j} g_j(z)$ and $g_j(x) > 0$ if and only if $x \in \{t \in \mathbb{R} : y_j = t \text{ for some } y \in S_X\}$. Therefore, X_j is a discrete RV with p.m.f. g_j. More generally, all marginal distributions of X are also discrete. The function g_j is usually referred to as the marginal p.m.f. of X_j.

Remark 1.293. Let $\emptyset \neq S \subset \mathbb{R}^p$ be a finite or countably infinite set and let $f : \mathbb{R}^p \to \mathbb{R}$ be such that

$$f(x) = 0, \forall x \in S^c, \quad f(x) > 0, \forall x \in S, \quad \sum_{x \in S} f(x) = 1.$$

Then f is the joint p.m.f. of some p-dimensional discrete random vector X with support S. We are not going to discuss the proof of this statement in this course.

Theorem 1.294. *Let $X = (X_1, X_2, \cdots, X_p)$ be a discrete random vector with joint DF F_X, joint p.m.f. f_X and support S_X. Let f_{X_j} denote the marginal p.m.f. of X_j. Then $X_1, X_2, \cdots, X_p$ are independent if and only if*

$$f_{X_1, X_2, \cdots, X_p}(x_1, x_2, \cdots, x_p) = \prod_{j=1}^{p} f_{X_j}(x_j), \forall x_1, x_2, \cdots, x_p \in \mathbb{R}.$$

In this case, we have $S_X = S_{X_1} \times S_{X_2} \times \cdots \times S_{X_p}$, where S_{X_j} denotes the support of X_j.

Proof. By Proposition 1.286, the RVs $X_1, X_2, \cdots, X_p$ are independent if and only if

$$F_{X_1, X_2, \cdots, X_p}(x_1, x_2, \cdots, x_p) = \prod_{j=1}^{p} F_{X_j}(x_j), \forall x_1, x_2, \cdots, x_p \in \mathbb{R}.$$

If the condition for the joint p.m.f. holds as per the statement above, then the above condition for the joint DF holds and hence the required independence follows.
The proof of the converse statement is left as an exercise in Problem set 7.

To prove the statement for the support, observe that

$$
\begin{aligned}
S_X &= \{x \in \mathbb{R}^p : f_X(x) > 0\} \\
&= \{x = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^p : \prod_{j=1}^{p} f_{X_j}(x_j) > 0\} \\
&= \{x = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^p : f_{X_j}(x_j) > 0, \forall j = 1, 2, \cdots, p\} \\
&= \prod_{j=1}^{p} \{x_j \in \mathbb{R} : f_{X_j}(x_j) > 0\} \\
&= S_{X_1} \times S_{X_2} \times \cdots \times S_{X_p}
\end{aligned}
$$

This completes the proof. □

Example 1.295. Given p.m.f.s $f_1, f_2, \cdots, f_p : \mathbb{R} \to [0,1]$ and corresponding support sets $S_1, S_2, \cdots, S_p$, consider the function $f : \mathbb{R}^p \to [0,1]$ defined by

$$f(x) := \prod_{j=1}^{p} f_j(x_j), \forall x = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^p.$$

Then the set $S = S_1 \times S_2 \times \cdots \times S_p \subset \mathbb{R}^p$ is also finite or countably infinite and

$$f(x) = 0, \forall x \in S^c, \quad f(x) > 0, \forall x \in S, \quad \sum_{x \in S} f(x) = 1.$$

By Remark 1.293, we have that f is the joint p.m.f. of a p-dimensional discrete random vector such that the component RVs are independent, by Theorem 1.294. Using this method, we can construct many examples of discrete random vectors.

Remark 1.296. Let $X = (X_1, X_2, \cdots, X_p)$ be a discrete random vector with joint p.m.f. f_X and support S_X. Then $X_1, X_2, \cdots, X_p$ are independent if and only if

$$f_{X_1, X_2, \cdots, X_p}(x_1, x_2, \cdots, x_p) = \prod_{j=1}^{p} g_j(x_j), \forall x_1, x_2, \cdots, x_p \in \mathbb{R}$$

for some functions $g_1, g_2, \cdots, g_p : \mathbb{R} \to [0, \infty)$ with $S_j := \{x \in \mathbb{R} : g_j(x) > 0\}$ being finite or countably infinite and $S_X = S_1 \times S_2 \times \cdots \times S_p$. In this case, the marginal p.m.fs f_{X_j} have the form $c_j g_j$, where the number c_j can be determined from the relation $c_j = \left(\sum_{x \in S_j} g_j(x)\right)^{-1}$.

Example 1.297. Let $Z = (X, Y)$ be a 2-dimensional discrete random vector with the joint p.m.f. of the form

$$f_Z(x, y) = \begin{cases} \alpha(x + y), & \text{if } x, y \in \{1, 2, 3, 4\} \\ 0, & \text{otherwise} \end{cases}$$

for some constant $\alpha \in \mathbb{R}$. For f_Z to take non-negative values, we must have $\alpha > 0$. Now, $\sum_{x,y \in \{1,2,3,4\}} \alpha(x + y) = 1$ simplifies to $80\alpha = 1$ and hence $\alpha = \frac{1}{80}$. Also note that for this value of α, f_Z takes non-negative values. The support of Z is $\{(x, y) : x, y \in \{1, 2, 3, 4\}\} = \{1, 2, 3, 4\} \times \{1, 2, 3, 4\}$. The support of X is $\{1, 2, 3, 4\}$ and the marginal p.m.f. f_X can now be computed as

$$f_X(x) = \begin{cases} \sum_{y \in \{1,2,3,4\}} \frac{1}{80}(x + y), & \text{if } x \in \{1, 2, 3, 4\} \\ 0, & \text{otherwise} \end{cases} = \begin{cases} \frac{1}{40}(2x + 5), & \text{if } x \in \{1, 2, 3, 4\} \\ 0, & \text{otherwise} \end{cases}$$

By the symmetry of $f_Z(x, y)$ in the variables x and y, we conclude that $X \overset{d}{=} Y$. Note that $f_Z(1, 1) = \frac{1}{40}$ and $f_X(1)f_Y(1) = \frac{49}{1600}$. Hence X and Y are not independent.

Example 1.298. Let $U = (X, Y, Z)$ be a 3-dimensional discrete random vector with the joint p.m.f. of the form

$$f_U(x, y, z) = \begin{cases} \alpha xyz, & \text{if } x = 1, y \in \{1, 2\}, z \in \{1, 2, 3\} \\ 0, & \text{otherwise} \end{cases}$$

for some constant $\alpha \in \mathbb{R}$. For f_U to take non-negative values, we must have $\alpha > 0$. Now, $\sum_{x=1, y \in \{1,2\}, z \in \{1,2,3\}} \alpha xyz = 1$ simplifies to $18\alpha = 1$ and hence $\alpha = \frac{1}{18}$. Also note that for this value of α, f_U takes non-negative values. The support of U is $\{(x, y, z) : x = 1, y \in \{1, 2\}, z \in \{1, 2, 3\}\} = \{1\} \times \{1, 2\} \times \{1, 2, 3\}$. The support of X is $\{1\}$ and the marginal p.m.f. f_X can now be computed as

$$f_X(x) = \begin{cases} \sum_{y\in\{1,2\},z\in\{1,2,3\}} \frac{1}{18}yz, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases} = \begin{cases} 1, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

as expected. Similar computation yields

$$f_Y(y) = \begin{cases} \frac{1}{3}, & \text{if } y = 1 \\ \frac{2}{3}, & \text{if } y = 2 \\ 0, & \text{otherwise} \end{cases} = \begin{cases} \frac{y}{3}, & \text{if } y \in \{1,2\} \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_Z(z) = \begin{cases} \frac{1}{6}, & \text{if } z = 1 \\ \frac{1}{3}, & \text{if } z = 2 \\ \frac{1}{2}, & \text{if } z = 3 \\ 0, & \text{otherwise} \end{cases} = \begin{cases} \frac{z}{6}, & \text{if } z \in \{1,2,3\} \\ 0, & \text{otherwise} \end{cases}$$

Observe that $f_{X,Y,Z}(x,y,z) = f_X(x)f_Y(y)f_Z(z), \forall x,y,z$ and hence the RVs X, Y, Z are independent.

Remark 1.299 (Conditional Distribution for discrete random vectors). Let $X = (X_1, X_2, \cdots, X_{p+q})$ be a discrete random vector with support S_X and joint p.m.f. f_X. Let $Y = (X_1, X_2, \cdots, X_p)$ and $Z = (X_{p+1}, X_{p+2}, \cdots, X_{p+q})$. Then Y and Z both are discrete random vectors. Let f_Y and S_Y denote the joint p.m.f. and support of Y, respectively. Let f_Z and S_Z denote the joint p.m.f. and support of Z, respectively. For $z \in S_Z$, consider the set

$$T_z := \{y \in \mathbb{R}^p : (y,z) \in S_X\}.$$

The conditional p.m.f. of Y given $Z = z \in S_Z$ is defined by

$$f_{Y|Z}(y \mid z) := \mathbb{P}(Y = y \mid Z = z) = \frac{\mathbb{P}(Y = y, Z = z)}{\mathbb{P}(Z = z)} = \begin{cases} \frac{f_X(y,z)}{f_Z(z)}, & \text{if } y \in T_z \\ 0, & \text{otherwise.} \end{cases}$$

By definition, $f_{Y|Z}(y \mid z) \geq 0, \forall y \in \mathbb{R}^p$ and $\sum_{y\in\mathbb{R}^p} f_{Y|Z}(y \mid z) = \sum_{y\in T_z} f_{Y|Z}(y \mid z) = 1$. Therefore, for every $z \in S_Z$, the function $y \in \mathbb{R}^p \mapsto f_{Y|Z}(y \mid z)$ is a joint p.m.f. with support T_z. We refer to the probability law/distribution described by this p.m.f. as the conditional distribution of Y given $Z = z \in S_Z$. The conditional DF of Y given $Z = z \in S_Z$ is given by

$$F_{Y|Z}(y \mid z) := \mathbb{P}(Y \leq y \mid Z = z) = \frac{\mathbb{P}(Y \leq y, Z = z)}{\mathbb{P}(Z = z)} = \sum_{\substack{t\leq y \\ t\in T_z}} \frac{f_X(t, z)}{f_Z(z)} = \sum_{\substack{t\leq y \\ t\in T_z}} f_{Y|Z}(t \mid z),$$

where $t \leq y$ refers to component-wise inequalities $t_j \leq y_j$ for all $j = 1, 2, \cdots, p$.

Note 1.300. For notational convenience, we have discussed the conditional distribution of first p component RVs with respect to the final q component RVs. However, as long as the $(p+q)$-dimensional joint distribution is known, we can discuss the conditional distribution of any of the k-component RVs with respect to the other $(p+q-k)$-component RVs.

Note 1.301. When values for some of the components RVs are given, the conditional distribution provides an updated probability distribution for the rest of the component RVs.

Note 1.302. Let (X, Y) be a 2-dimensional discrete random vector such that X and Y are independent. Then $f_{X,Y}(x, y) = f_X(x)f_Y(y), \forall x, y \in \mathbb{R}$. Then

$$f_{Y|X}(y \mid x) = f_Y(y), \forall x \in S_X, y \in S_Y.$$

This statement can be generalized to higher dimensions with appropriate changes in the notation.

Example 1.303. In Example 1.297, we have, for fixed $x \in \{1, 2, 3, 4\}$,

$$f_{Y|X}(y \mid x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)}, & \text{if } y \in \{1,2,3,4\} \\ 0, & \text{otherwise.} \end{cases} = \begin{cases} \frac{x+y}{2(2x+5)}, & \text{if } y \in \{1,2,3,4\} \\ 0, & \text{otherwise.} \end{cases}$$

Definition 1.304 (Continuous Random Vector and its Joint Probability Density Function (Joint p.d.f.)). A random vector $X = (X_1, X_2, \cdots, X_p)$ is said to be a continuous random vector if there exists an integrable function $f : \mathbb{R}^p \to [0, \infty)$ such that

$$F_X(x) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_p \leq x_p)$$

$$= \int_{t_1=-\infty}^{x_1} \int_{t_2=-\infty}^{x_2} \cdots \int_{t_p=-\infty}^{x_p} f(t_1, t_2, \cdots, t_p)\, dt_p dt_{p-1} \cdots dt_2 dt_1, \forall x = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^p.$$

The function f is called the joint probability density function (joint p.d.f.) of X.

Remark 1.305. Let X be a continuous random vector with joint DF F_X and joint p.d.f. f_X. Then we have the following observations.

(a) F_X is jointly continuous in all co-ordinates.

(b) $\mathbb{P}(X = x) = 0, \forall x \in \mathbb{R}^p$. More generally, if $A \subset \mathbb{R}^p$ is finite or countably infinite, then by the finite/countable additivity of $\mathbb{P}_X$, we have

$$\mathbb{P}(X \in A) = \mathbb{P}_X(A) = \sum_{x \in A} \mathbb{P}_X(\{x\}) = \sum_{x \in A} \mathbb{P}(X = x) = 0.$$

(c) By definition, we have $f_X(x) \geq 0, \forall x \in \mathbb{R}^p$ and

$$\begin{aligned}
1 &= \lim_{x_j \to \infty \, \forall j} F_X(x_1, x_2, \cdots, x_p) \\
&= \lim_{x_j \to \infty \, \forall j} \int_{t_1=-\infty}^{x_1} \int_{t_2=-\infty}^{x_2} \cdots \int_{t_p=-\infty}^{x_p} f(t_1, t_2, \cdots, t_p)\, dt_p dt_{p-1} \cdots dt_2 dt_1 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(t_1, t_2, \cdots, t_p)\, dt_p dt_{p-1} \cdots dt_2 dt_1.
\end{aligned}$$

(d) Suppose that the joint p.d.f. f_X of a p-dimensional random vector X is piecewise continuous. Then by the Fundamental Theorem of Calculus (specifically multivariable Calculus), we have

$$f_X(x_1, x_2, \cdots, x_p) = \frac{\partial^p}{\partial x_1 \partial x_2 \cdots \partial x_p} F_X(x_1, x_2, \cdots, x_p),$$

wherever the partial derivative on the right hand side exists.

(e) If X is a p-dimensional random vector such that its joint DF F_X is continuous on $\mathbb{R}^p$ and such that the partial derivative $\frac{\partial^p}{\partial x_1 \partial x_2 \cdots \partial x_p} F_X$ exists everywhere except possibly on a countable number of curves on $\mathbb{R}^p$. Let $A \subset \mathbb{R}^p$ denote the set of all points on such curves. Then X is a continuous random vector with the joint p.d.f.

$$f_X(x) = \begin{cases} \frac{\partial^p}{\partial x_1 \partial x_2 \cdots \partial x_p} F_X(x), \text{ if } x = (x_1, x_2, \cdots, x_p) \in A^c, \\ 0, \text{ if } x = (x_1, x_2, \cdots, x_p) \in A. \end{cases}$$

(f) The joint p.d.f. of a continuous random vector is not unique. As in the case of continuous RVs, the joint p.d.f. is determined uniquely upto sets of 'volume 0'. Here, we also get versions of the joint p.d.f..

(g) For $A \subset \mathbb{R}^p$, we have

$$\begin{aligned}\mathbb{P}(X \in A) &= \iiint_A f_X(t_1, t_2, \cdots, t_p)\, dt_p dt_{p-1} \cdots dt_2 dt_1\\ &= \iiint_{\mathbb{R}^p} f_X(t_1, t_2, \cdots, t_p) 1_A(t_1, t_2, \cdots, t_p)\, dt_p dt_{p-1} \cdots dt_2 dt_1,\end{aligned}$$

provided the integral can be defined. We do not prove this statement in this course.

(h) For any $j \in \{1, 2, \cdots, p\}$, for $x_j \in \mathbb{R}$

$$\begin{aligned}F_{X_j}(x_j) &= \mathbb{P}(X_j \in (-\infty, x_j])\\ &= \mathbb{P}(X_1 \in \mathbb{R}, \cdots, X_{j-1} \in \mathbb{R}, X_j \in (-\infty, x_j], X_{j+1} \in \mathbb{R}, \cdots, X_p \in \mathbb{R})\\ &= \mathbb{P}(X \in \mathbb{R} \times \cdots \times \mathbb{R} \times (-\infty, x_j] \times \mathbb{R} \times \cdots \times \mathbb{R})\\ &= \int_{t_1=-\infty}^{\infty} \cdots \int_{t_{j-1}=-\infty}^{\infty} \int_{t_j=-\infty}^{x_j} \int_{t_{j+1}=-\infty}^{\infty} \cdots \int_{t_p=-\infty}^{\infty} f_X(t_1, t_2, \cdots, t_p)\, dt_p dt_{p-1} \cdots dt_2 dt_1.\end{aligned}$$

Consider $g_j : \mathbb{R} \to \mathbb{R}$ defined by

$$g_j(t_j) := \int_{t_1=-\infty}^{\infty} \cdots \int_{t_{j-1}=-\infty}^{\infty} \int_{t_{j+1}=-\infty}^{\infty} \cdots \int_{t_p=-\infty}^{\infty} f_X(t_1, t_2, \cdots, t_p)\, dt_p \cdots dt_{j-1} dt_{j+1} \cdots dt_2 dt_1.$$

It is immediate that g_j satisfies the properties of a p.d.f. and $F_{X_j}(x_j) = \int_{t_j=-\infty}^{x_j} g_j(t_j)\, dt_j$. Therefore, X_j is a continuous RV with p.d.f. g_j. More generally, all marginal distributions of X are also continuous and can be obtained by integrating out the unnecessary co-ordinates. The function g_j is usually referred to as the marginal p.d.f. of X_j.

Remark 1.306. Let $f : \mathbb{R}^p \to [0, \infty)$ be an integrable function with

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(t_1, t_2, \cdots, t_p)\, dt_p dt_{p-1} \cdots dt_2 dt_1 = 1.$$

Then f is the joint p.d.f. of some p-dimensional continuous random vector X. We are not going to discuss the proof of this statement in this course.

We can identify the independence of the component RVs for a continuous random vector via the joint p.d.f.. The proof is similar to Theorem 1.294 and is skipped for brevity.

Theorem 1.307. *Let $X = (X_1, X_2, \cdots, X_p)$ be a continuous random vector with joint DF F_X, joint p.d.f. f_X. Let f_{X_j} denote the marginal p.d.f. of X_j. Then $X_1, X_2, \cdots, X_p$ are independent if and only if*

$$f_{X_1,X_2,\cdots,X_p}(x_1, x_2, \cdots, x_p) = \prod_{j=1}^{p} f_{X_j}(x_j), \forall x_1, x_2, \cdots, x_p \in \mathbb{R}.$$

Example 1.308. Given p.d.f.s $f_1, f_2, \cdots, f_p : \mathbb{R} \to [0, \infty)$, consider the function $f : \mathbb{R}^p \to [0, \infty)$ defined by

$$f(x) := \prod_{j=1}^{p} f_j(x_j), \forall x = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^p.$$

Then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(t_1, t_2, \cdots, t_p)\, dt_p dt_{p-1} \cdots dt_2 dt_1 = 1.$$

By Remark 1.306, we have that f is the joint p.d.f. of a p-dimensional continuous random vector such that the component RVs are independent, by Theorem 1.307. Using this method, we can construct many examples of continuous random vectors.

Remark 1.309. Let $X = (X_1, X_2, \cdots, X_p)$ be a continuous random vector with joint p.d.f. f_X. Then $X_1, X_2, \cdots, X_p$ are independent if and only if

$$f_{X_1,X_2,\cdots,X_p}(x_1, x_2, \cdots, x_p) = \prod_{j=1}^{p} g_j(x_j), \forall x_1, x_2, \cdots, x_p \in \mathbb{R}$$

for some integrable functions $g_1, g_2, \cdots, g_p : \mathbb{R} \to [0, \infty)$. In this case, the marginal p.d.fs f_{X_j} have the form $c_j g_j$, where the number c_j can be determined from the relation $c_j = \left(\int_{-\infty}^{\infty} g_j(x)\, dx\right)^{-1}$.

Example 1.310. Let $Z = (X, Y)$ be a 2-dimensional continuous random vector with the joint p.d.f. of the form

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

for some constant $\alpha \in \mathbb{R}$. For f_Z to take non-negative values, we must have $\alpha > 0$. Now,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_Z(x,y)\,dxdy = \int_{y=0}^{1}\int_{x=0}^{y} \alpha xy\,dxdy = \int_{y=0}^{1} \alpha\frac{y^3}{2}\,dy = \frac{\alpha}{8}.$$

For f_Z to be a joint p.d.f., we need $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_Z(x,y)\,dxdy = 1$ and hence $\alpha = 8 > 0$. Also note that for this value of α, f_Z takes non-negative values. The marginal p.d.f. f_X of X can now be computed as follows.

$$f_X(x) = \int_{-\infty}^{\infty} f_Z(x,y)\,dy = \begin{cases} \int_{y=x}^{1} 8xy\,dy, \text{ if } x \in (0,1) \\ 0, \text{ otherwise} \end{cases} = \begin{cases} 4x[1-x^2], \text{ if } x \in (0,1) \\ 0, \text{ otherwise} \end{cases}.$$

The marginal p.d.f. f_Y of Y follows by a similar computation.

$$f_Y(y) = \int_{-\infty}^{\infty} f_Z(x,y)\,dx = \begin{cases} \int_{x=0}^{y} 8xy\,dx, \text{ if } y \in (0,1) \\ 0, \text{ otherwise} \end{cases} = \begin{cases} 4y^3, \text{ if } y \in (0,1) \\ 0, \text{ otherwise} \end{cases}.$$

Observe that $f_Z(\frac{1}{2},\frac{1}{2}) = 0$ and $f_X(\frac{1}{2})f_Y(\frac{1}{2}) = \frac{3}{2} \times \frac{1}{2} = \frac{3}{4}$. Hence X and Y are not independent.

Example 1.311. Let $U = (X,Y,Z)$ be a 3-dimensional continuous random vector with the joint p.d.f. of the form

$$f_U(x,y,z) = \begin{cases} \alpha xyz, \text{ if } x,y,z \in (0,1) \\ 0, \text{ otherwise} \end{cases}$$

for some constant $\alpha \in \mathbb{R}$. For f_Z to take non-negative values, we must have $\alpha > 0$. Now,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_U(x,y,z)\,dxdydz = \int_{x=0}^{1}\int_{y=0}^{1}\int_{z=0}^{1} \alpha xyz\,dxdydz = \frac{\alpha}{8}.$$

For f_U to be a joint p.d.f., we need $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_U(x,y,z)\,dxdydz = 1$ and hence $\alpha = 8 > 0$. Also note that for this value of α, f_U takes non-negative values. The marginal p.d.f. f_X of X can now be computed as follows.

$$f_X(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_U(x,y,z)\,dydz = \begin{cases} \int_{z=0}^{1}\int_{y=0}^{1} 8xyz\,dydz, \text{ if } x \in (0,1) \\ 0, \text{ otherwise} \end{cases} = \begin{cases} 2x, \text{ if } x \in (0,1) \\ 0, \text{ otherwise} \end{cases}.$$

By the symmetry of $f_U(x,y,z)$ in the variables x,y and z, we conclude that $X \overset{d}{=} Y \overset{d}{=} Z$. Observe that $f_{X,Y,Z}(x,y,z) = f_X(x)f_Y(y)f_Z(z), \forall x,y,z$ and hence the RVs X,Y,Z are independent.

Note 1.312. There are random vectors which are neither discrete nor continuous. We do not discuss such examples in this course.

Remark 1.313 (Conditional Distribution for continuous random vectors)*.* We now discuss an analogue of conditional distributions as discussed in Remark 1.299 for discrete random vectors. To avoid notational complexity, we work in dimension 2. Let (X,Y) be a 2-dimensional continuous random vector with joint DF $F_{X,Y}$ and joint p.d.f. $f_{X,Y}$. Let f_X and f_Y denote the marginal p.d.fs of X and Y respectively. Since $\mathbb{P}(X = x) = 0, \forall x \in \mathbb{R}$, expressions of the form $\mathbb{P}(Y \in A \mid X = x)$ are not defined for $A \subset \mathbb{R}$. We consider $x \in \mathbb{R}$ such that $f_X(x) > 0$ and look at the following computation. For $y \in \mathbb{R}$,

$$\begin{aligned}
\lim_{h \downarrow 0} \mathbb{P}(Y \leq y \mid x - h < X \leq x) &= \lim_{h \downarrow 0} \frac{\mathbb{P}(Y \leq y, x - h < X \leq x)}{\mathbb{P}(x - h < X \leq x)} \\
&= \lim_{h \downarrow 0} \frac{\int_{x-h}^{x} \int_{-\infty}^{y} f_{X,Y}(t,s)\, dsdt}{\int_{x-h}^{x} f_X(t)\, dt} \\
&= \lim_{h \downarrow 0} \frac{\frac{1}{h}\int_{x-h}^{x} \int_{-\infty}^{y} f_{X,Y}(t,s)\, dsdt}{\frac{1}{h}\int_{x-h}^{x} f_X(t)\, dt} \\
&= \frac{\int_{-\infty}^{y} f_{X,Y}(x,s)\, ds}{f_X(x)} \\
&= \int_{-\infty}^{y} \frac{f_{X,Y}(x,s)}{f_X(x)}\, ds
\end{aligned}$$

Here, we have assumed continuity of the p.d.fs. Motivated by the above computation, we define the conditional DF of Y given $X = x$ (provided $f_X(x) > 0$) by

$$F_{Y|X}(y \mid x) := \lim_{h \downarrow 0} \mathbb{P}(Y \leq y \mid x - h < X \leq x),\ y \in \mathbb{R}$$

and the conditional p.d.f. of Y given $X = x$ (provided $f_X(x) > 0$) by

$$f_{Y|X}(y \mid x) := \frac{f_{X,Y}(x,y)}{f_X(x)},\ y \in \mathbb{R}.$$

These calculations generalizes to the higher dimensions as follows. Let $X = (X_1, X_2, \cdots, X_{p+q})$ be a continuous random vector with joint p.d.f. f_X. Let $Y = (X_1, X_2, \cdots, X_p)$ and $Z = (X_{p+1}, X_{p+2}, \cdots, X_{p+q})$. If $z \in \mathbb{R}^q$ be such that $f_Z(z) > 0$, then we define the conditional DF of Y given $Z = z$ by

$$F_{Y|Z}(y \mid z) := \lim_{\substack{h_j \downarrow 0 \\ j=p+1,p+2,\cdots,p+q}} \mathbb{P}(X_1 \le y_1, \cdots, X_p \le y_p \mid x_j - h_j < X_j \le x_j, \forall j),\, y \in \mathbb{R}^p$$

and the conditional p.d.f. of Y given $Z = z$ by

$$f_{Y|Z}(y \mid z) := \frac{f_{Y,Z}(y,z)}{f_Z(z)},\, y \in \mathbb{R}^p.$$

Note 1.314. For notational convenience, we have discussed the conditional distribution of first p component RVs with respect to the final q component RVs. However, as long as the $(p+q)$-dimensional joint distribution is known, we can discuss the conditional distribution of any of the k-component RVs with respect to the other $(p+q-k)$-component RVs.

Note 1.315. Let (X, Y) be a 2-dimensional continuous random vector such that X and Y are independent. Then $f_{X,Y}(x,y) = f_X(x)f_Y(y), \forall x, y \in \mathbb{R}$. Then

$$f_{Y|X}(y \mid x) = f_Y(y), \forall y \in \mathbb{R},$$

provided $f_X(x) > 0$. This statement can be generalized to higher dimensions with appropriate changes in the notation.

Example 1.316. In Example 1.310, we have, for fixed $x \in (0,1)$,

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \begin{cases} \frac{2xy}{x(1-x^2)}, & \text{if } y \in (x,1) \\ 0, & \text{otherwise.} \end{cases}$$

Earlier in Week 5, we have discussed about the distribution of functions of RVs. We now generalize the same concept for random vectors.

Remark 1.317. Let $X = (X_1, \dots, X_p)$ be a p-dimensional discrete/continuous random vector with joint p.m.f./p.d.f. f_X. We are interested in the distribution of $Y = h(X)$ for functions $h : \mathbb{R}^p \to$

$\mathbb{R}^q$. Here, $Y = (Y_1, \ldots, Y_q)$ is a q-dimensional random vector with $Y_j = h_j(X_1, \ldots, X_p)$, where $h_j : \mathbb{R}^p \to \mathbb{R}, j = 1, 2, \cdots, q$ denotes the component functions of h. The distribution of Y is uniquely determined as soon as we are able to compute the joint DF F_Y of Y. Note that

$$F_Y(y_1, \cdots, y_q) = \mathbb{P}(Y_1 \leq y_1, \cdots, Y_q \leq y_q) = \mathbb{P}(h_1(X) \leq y_1, \cdots, h_q(X) \leq y_q), \forall (y_1, \cdots, y_q) \in \mathbb{R}^q.$$

Once the joint DF F_Y is known, the joint p.m.f./p.d.f. of Y can then be deduced by standard techniques.

Example 1.318. Let $X_1 \sim Uniform(0,1)$ and $X_2 \sim Uniform(0,1)$ be independent RVs. Suppose we are interested in the distribution of $Y = X_1 + X_2$. By independence of X_1 and X_2, the joint p.d.f. (X_1, X_2) is given by

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

Consider the function $h : \mathbb{R}^2 \to \mathbb{R}$ defined by $h(x_1, x_2) := x_1 + x_2, \forall (x_1, x_2) \in \mathbb{R}^2$. Then $Y = h(X_1, X_2)$. Now, for $y \in \mathbb{R}$

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(h(X_1, X_2) \leq y) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{(-\infty,y]}(h(x_1, x_2)) f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 \\ &= \int_0^1 \int_0^1 1_{(-\infty,y]}(x_1 + x_2)\, dx_1 dx_2 \\ &= \begin{cases} 0, \text{ if } y < 0, \\ \int_{x_1=0}^{y} \int_{x_2=0}^{y-x_1} dx_2 dx_1, \text{ if } 0 \leq y < 1, \\ 1 - \frac{1}{2} \times (2-y) \times (2-y), \text{ if } 1 \leq y < 2, \\ 1, \text{ if } y \geq 2 \end{cases} \end{aligned}$$

$$= \begin{cases} 0, \text{ if } y < 0, \\ \frac{y^2}{2}, \text{ if } 0 \leq y < 1, \\ \frac{4y-y^2-2}{2}, \text{ if } 1 \leq y < 2, \\ 1, \text{ if } y \geq 2 \end{cases}$$

Here, F_Y is differentiable everywhere except possibly at the points $0, 1, 2$ and

$$F'_Y(y) = \begin{cases} y, \text{ if } y \in (0,1), \\ 2-y, \text{ if } y \in (1,2), \\ 0, \text{ otherwise.} \end{cases}$$

Observe that $\int_{-\infty}^{\infty} F'_Y(y)\,dy = 1$ and the derivative is non-negative. Hence, Y is a continuous RV with the p.d.f. given by F'_Y.

As done in the case of RVs, in the setting of Remark 1.317, we consider the computation of the joint p.m.f./p.d.f. of Y directly, instead of computing the joint DF F_Y first. The next result is a direct generalization of Theorem 1.164 and we skip the proof for brevity.

Theorem 1.319 (Change of Variables for Discrete random vectors)**.** *Let $X = (X_1, \ldots, X_p)$ be a p-dimensional discrete random vector with joint p.m.f. f_X and support S_X. Let $h = (h_1, \cdots, h_q) : \mathbb{R}^p \to \mathbb{R}^q$ be a function and let $Y = (Y_1, \cdots, Y_q) = h(X) = (h_1(X), \cdots, h_q(X))$. Then Y is a discrete random vector with support*

$$S_Y = h(S_X) = \{h(x) : x \in S_X\},$$

joint p.m.f.

$$f_Y(y) = \begin{cases} \sum\limits_{\substack{x \in S_X \\ h(x)=y}} f_X(x), \ \textit{if } y \in S_Y, \\ 0, \ \textit{otherwise} \end{cases}$$

and joint DF

$$F_Y(y) = \sum_{\substack{x \in S_X \\ h(x) \leq y}} f_X(x), \forall y \in \mathbb{R}^q.$$

Example 1.320. Fix $p \in (0,1)$ and let $n_1, \cdots, n_q$ be positive integers. Let $X_1, \cdots, X_q$ be independent RVs with $X_i \sim Binomial(n_i, p), i = 1, \cdots, q$. Here, the with p.m.f.s are given by

$$f_{X_i}(x_i) = \begin{cases} \binom{n_i}{x} p^x (1-p)^{n_i - x}, \forall x \in \{0, 1, \cdots, n_i\}, \\ 0, \text{ otherwise} \end{cases}$$

for $i = 1, \cdots, q$. Using independence, the joint p.m.f. is given by

$$f_X(x_1, \cdots, x_q) = \begin{cases} \prod_{i=1}^q \binom{n_i}{x_i} p^{\sum_{i=1}^q x_i} (1-p)^{n - \sum_{i=1}^q x_i}, \forall (x_1, \cdots, x_q) \in \prod_{i=1}^q \{0, 1, \cdots, n_i\}, \\ 0, \text{ otherwise} \end{cases}$$

where $n = n_1 + \cdots + n_q$. Consider $Y = X_1 + \cdots + X_q$. Now, if $y \notin \{0, 1, \cdots, n\}$, $f_Y(y) = \mathbb{P}(X_1 + \cdots + X_q = y) = 0$ and if $y \in \{0, 1, \cdots, n\}$, then

$$\begin{aligned} f_Y(y) &= \mathbb{P}(X_1 + \cdots + X_q = y) \\ &= \sum_{\substack{(x_1, \cdots, x_q) \in \prod_{i=1}^q \{0, 1, \cdots, n_i\} \\ x_1 + \cdots + x_q = y}} f_X(x_1, \cdots, x_q) \\ &= p^y (1-p)^{n-y} \sum_{\substack{(x_1, \cdots, x_q) \in \prod_{i=1}^q \{0, 1, \cdots, n_i\} \\ x_1 + \cdots + x_q = y}} \prod_{i=1}^q \binom{n_i}{x_i} \\ &= \binom{n}{y} p^y (1-p)^{n-y}. \end{aligned}$$

Therefore, $Y = X_1 + \cdots + X_q \sim Binomial(n, p)$ with $n = n_1 + \cdots + n_q$.

Remark 1.321. We had earlier mentioned that $Bernoulli(p)$ distribution is the same as $Binomial(1, p)$ distribution. Using the above computation, we can identify a $Binomial(n, p)$ RV as a sum of n independent RVs each having distribution $Bernoulli(p)$. We shall come back to this observation in later lectures.

For continuous random vectors, we have the following generalization of Theorem 1.172. Proof of this result is being skipped.

Theorem 1.322. *Let $X = (X_1, \ldots, X_p)$ be a p-dimensional continuous random vector with joint p.d.f. f_X. Suppose that $\{x \in \mathbb{R}^p : f_X(x) > 0\}$ can be written as a disjoint union $\cup_{i=1}^k S_i$ of open sets in $\mathbb{R}^p$.*

Let $h^j : \mathbb{R}^p \to \mathbb{R}, j = 1, \cdots, p$ be functions such that $h = (h^1, \cdots, h^p) : S_i \to \mathbb{R}^p$ is one-to-one with inverse $h_i^{-1} = ((h_i^1)^{-1}, \cdots, (h_i^p)^{-1})$ for each $i = 1, \cdots, k$. Moreover, assume that $(h_i^j)^{-1}, i = 1, 2, \cdots, k; j = 1, \cdots, p$ have continuous partial derivatives and the Jacobian determinant of the transformation

$$J_i := \begin{vmatrix} \frac{\partial (h_i^1)^{-1}}{\partial y_1}(t) & \cdots & \frac{\partial (h_i^1)^{-1}}{\partial y_p}(y) \\ \vdots & \vdots & \vdots \\ \frac{\partial (h_i^p)^{-1}}{\partial y_1}(y) & \cdots & \frac{\partial (h_i^p)^{-1}}{\partial y_p}(y) \end{vmatrix} \neq 0, \forall i = 1, \cdots, k.$$

Then the p-dimensional random vector $Y = (Y_1, \cdots, Y_p) = h(X) = (h^1(X), \cdots, h^p(X))$ is a continuous with joint p.d.f.

$$f_Y(y) = \sum_{i=1}^k f_X((h_i^1)^{-1}(y), \cdots, (h_i^p)^{-1}(y))\, |J_i|\, 1_{h(S_i)}(y).$$

Example 1.323. Fix $\lambda > 0$. Let $X_1 \sim Exponential(\lambda)$ and $X_2 \sim Exponential(\lambda)$ be independent RVs defined on the same probability space. The joint distribution of (X_1, X_2) is given by the joint p.d.f.

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = \begin{cases} \frac{1}{\lambda^2} \exp(-\frac{x_1+x_2}{\lambda}), & \text{if } x_1 > 0, x_2 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Consider the function

$$h(x_1, x_2) = \begin{cases} (x_1 + x_2, \frac{x_1}{x_1+x_2}), \forall x_1 > 0, x_2 > 0, \\ 0, \text{otherwise.} \end{cases}$$

Here, $\{(x_1, x_2) \in \mathbb{R}^2 : f_{X_1,X_2}(x_1, x_2) > 0\} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > 0, x_2 > 0\}$ and $h : \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > 0, x_2 > 0\} \to \mathbb{R}^2$ is one-to-one with range $(0, \infty) \times (0, 1)$. The inverse function is

$h^{-1}(y_1, y_2) = (y_1y_2, y_1(1-y_2))$ for $(y_1, y_2) \in (0,\infty) \times (0,1)$ with Jacobian determinant given by

$$J(y_1, y_2) = \begin{vmatrix} y_2 & y_1 \\ 1-y_2 & -y_1 \end{vmatrix} = -y_1.$$

Now, $Y = (Y_1, Y_2) = h(X_1, X_2) = (X_1 + X_2, \frac{X_1}{X_1+X_2})$ has the joint p.d.f. given by

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= \begin{cases} f_{X_1,X_2}(y_1y_2, y_1(1-y_2))|J(y_1,y_2)|, \text{ if } y_1 > 0, y_2 \in (0,1) \\ 0, \text{ otherwise.} \end{cases} \\ &= \begin{cases} \frac{1}{\lambda^2} y_1 \exp\left(-\frac{y_1}{\lambda}\right), \text{ if } y_1 > 0, y_2 \in (0,1) \\ 0, \text{ otherwise.} \end{cases} \end{aligned}$$

Now, we compute the marginal distributions. The marginal p.d.f. f_{Y_1} is given by

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 = \begin{cases} \frac{1}{\lambda^2} y_1 \exp\left(-\frac{y_1}{\lambda}\right), \text{ if } y_1 > 0 \\ 0, \text{ otherwise} \end{cases}$$

and the marginal p.d.f. f_{Y_2} is given by

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_1 = \begin{cases} 1, \text{ if } y_2 \in (0,1) \\ 0, \text{ otherwise} \end{cases}.$$

Therefore $Y_1 = X_1 + X_2 \sim Gamma(2, \lambda)$ and $Y_2 = \frac{X_1}{X_1+X_2} \sim Uniform(0,1)$. Moreover,

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2), \forall (y_1, y_2) \in \mathbb{R}^2$$

and hence Y_1 and Y_2 are independent.

Remark 1.324. We had earlier mentioned that $Exponential(\lambda)$ distribution is the same as $Gamma(1, \lambda)$ distribution. Using the above computation, we can identify a $Gamma(2, \lambda)$ RV as a sum of two independent RVs each having distribution $Gamma(1, \lambda)$. A more general property in this regard is mentioned in practice problem set 8.

We now consider expectations for random vectors and for functions of random vectors. The concepts are same as discussed in the case of RVs.

Definition 1.325 (Expectation/Mean/Expected Value for functions of Random Vectors)**.** Let $X = (X_1, X_2, \cdots, X_p)$ be a p-dimensional discrete/continuous random vector with joint p.m.f./p.d.f. f_X. Let $h : \mathbb{R}^p \to \mathbb{R}$ be a function. Then $h(X)$ is an one-dimensional random vectors, i.e. an RV. We say that the expectation of $h(X)$, denoted by $\mathbb{E}h(X)$, is defined as the quantity

$$\mathbb{E}h(X) := \begin{cases} \sum_{x \in S_X} h(x) f_X(x), & \text{if } \sum_{x \in S_X} |h(x)| f_X(x) < \infty \text{ for discrete } X, \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x) f_X(x)\, dx, & \text{if } \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |h(x)| f_X(x)\, dx < \infty \text{ for continuous } X. \end{cases}$$

In the discrete case, S_X denotes the support of X.

Remark 1.326. If the sum or the integral above converges absolutely, we say that the expectation $\mathbb{E}h(X)$ exists or equivalently, $\mathbb{E}h(X)$ is finite. Otherwise, we shall say that the expectation $\mathbb{E}h(X)$ does not exist.

The following results is a generalization of Proposition 1.190. We skip the proof for brevity.

Proposition 1.327. *(a) Let $X = (X_1, X_2, \cdots, X_p)$ be a discrete random vector with joint p.m.f. f_X and support S_X and let $h : \mathbb{R}^p \to \mathbb{R}$ be a function. Consider the discrete RV $Y := h(X)$ with p.m.f. f_Y and support S_Y. Then $\mathbb{E}Y$ exists if and only if $\sum_{y \in S_Y} |y| f_Y(y) < \infty$ and in this case,*

$$\mathbb{E}Y = \mathbb{E}h(X) = \sum_{x \in S_X} h(x) f_X(x) = \sum_{y \in S_Y} y f_Y(y).$$

(b) Let $X = (X_1, X_2, \cdots, X_p)$ be a continuous random vector with joint p.d.f. f_X. Let $h : \mathbb{R}^p \to \mathbb{R}$ be a function such that the RV $Y := h(X)$ is continuous with p.d.f. f_Y. Then $\mathbb{E}Y$ exists if and only if $\int_{-\infty}^{\infty} |y| f_Y(y)\, dy < \infty$ and in this case,

$$\mathbb{E}Y = \mathbb{E}h(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x) f_X(x)\, dx = \int_{-\infty}^{\infty} y f_Y(y)\, dy.$$

Note 1.328. As considered for the case of RVs, by choosing different functions $h : \mathbb{R}^p \to \mathbb{R}$, we obtain several quantities of interest of the form $\mathbb{E}h(X)$ for a p-dimensional random vector X.

Definition 1.329 (Some special expectations for Random Vectors)**.** Let $X = (X_1, X_2, \cdots, X_p)$ be a p-dimensional discrete/continuous random vector.

(a) (Joint Moments) For non-negative integers $k_1, \ldots, k_p$, let $h(x) := x_1^{k_1} \cdots x_p^{k_p}, \forall x \in \mathbb{R}^p$. Then,

$$\mu'_{k_1,\ldots,k_p} := \mathbb{E}\left(X_1^{k_1} \cdots X_p^{k_p}\right)$$

is called a joint moment of order $k_1 + \cdots + k_p$ of X, provided it exists.

(b) (Joint Central Moments) For non-negative integers $k_1, \ldots, k_p$, let

$$h(x) := (x_1 - \mathbb{E}(X_1))^{k_1} \cdots (x_p - \mathbb{E}(X_p))^{k_p}, \forall x \in \mathbb{R}^p.$$

Then

$$\mu_{k_1,\ldots,k_p} := \mathbb{E}\left((X_1 - \mathbb{E}(X_1))^{k_1} \cdots (X_p - \mathbb{E}(X_p))^{k_p}\right)$$

is called a joint central moment of order $k_1 + \cdots + k_p$ of X, provided it exists.

(c) (Covariance) Fix $i, j = 1, \ldots, p$. Let $h(x) := (x_i - \mathbb{E}(X_i))(x_j - \mathbb{E}(X_j)), \forall x = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^p$. Then, $\mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))]$ is called the covariance between X_i and X_j, provided it exists. We shall denote this quantity by $Cov(X_i, X_j)$.

(d) (Joint Moment Generating Function, or simply, Joint MGF) We define

$$A := \left\{t = (t_1, t_2, \ldots, t_p) \in \mathbb{R}^p : \mathbb{E}\left(e^{\sum_{i=1}^p t_i X_i}\right) < \infty\right\},$$

and consider the function $M_X : A \to \mathbb{R}$ defined by

$$M_X(t) = \mathbb{E}\left(e^{\sum_{i=1}^p t_i X_i}\right), \quad \forall t = (t_1, t_2, \ldots, t_p) \in A.$$

The function M_X is called the joint moment generating function (joint MGF) of the random vector X. Note that $t = (0, 0, \cdots, 0) \in \mathbb{R}^p$ yields $M_X(t) = 1$ and hence $(0, 0, \cdots, 0) \in A$.

Remark 1.330*.* We now list some properties of the above quantities. The properties are being stated under the assumption that the expectations involved exist. Let $X = (X_1, X_2, \cdots, X_p)$ be a p-dimensional discrete/continuous random vector.

(a) Let $a_1, \dots, a_p$ be real constants. Then, $\mathbb{E}\left(\sum_{i=1}^{p} a_i X_i\right) = \sum_{i=1}^{p} a_i \mathbb{E}X_i$. To see this for discrete X, observe that

$$\mathbb{E}\left(\sum_{i=1}^{p} a_i X_i\right) = \sum_{x \in S_X} \sum_{i=1}^{p} a_i x_i f_X(x) = \sum_{i=1}^{p} \sum_{x \in S_X} a_i x_i f_X(x) = \sum_{i=1}^{p} a_i \mathbb{E}X_i.$$

The interchange of the order of summation is allowed due to absolute convergence of the series involved. The proof for continuous X is similar.

(b) $Cov(X_i, X_j) = Cov(X_j, X_i)$, for all $i, j = 1, \dots, p$.

(c) $Cov(X_i, X_i) = Var(X_i)$, for all $i = 1, \dots, p$.

(d) For all $i, j = 1, \dots, p$, we have

$$\begin{aligned} Cov(X_i, X_j) &= \mathbb{E}\left[X_i X_j - X_i(\mathbb{E}X_j) - X_j(\mathbb{E}X_i) + (\mathbb{E}X_i)(\mathbb{E}X_j)\right] \\ &= \mathbb{E}\left(X_i X_j\right) - \mathbb{E}\left(X_i\right)\mathbb{E}\left(X_j\right) \end{aligned}$$

(e) Let $X_1, X_2, \dots, X_p, Y_1, Y_2, \dots, Y_q$ be RVs, and let $a_1, \dots, a_p, b_1, \dots, b_q$ be real constants. Then,

$$Cov\left(\sum_{i=1}^{p} a_i X_i, \sum_{j=1}^{q} b_j Y_j\right) = \sum_{i=1}^{p} \sum_{j=1}^{q} a_i b_j Cov\left(X_i, Y_j\right).$$

In particular,

$$\begin{aligned} Var\left(\sum_{i=1}^{p} a_i X_i\right) &= \sum_{i=1}^{p} a_i^2 Var\left(X_i\right) + \sum_{i=1}^{p} \sum_{\substack{j=1 \\ j \neq i}}^{p} a_i a_j Cov\left(X_i, X_j\right) \\ &= \sum_{i=1}^{p} a_i^2 Var\left(X_i\right) + 2 \sum_{1 \le i < j \le p} a_i a_j Cov\left(X_i, X_j\right). \end{aligned}$$

(f) Let $X_1, X_2, \cdots, X_p$ be independent and let $h_1, h_2, \cdots, h_p : \mathbb{R} \to \mathbb{R}$ be functions. Then

$$\mathbb{E}\left(\prod_{i=1}^{p} h_i(X_i)\right) = \prod_{i=1}^{p} \mathbb{E}h_i(X_i).$$

For simplicity, we discuss the proof when $p = 2$ and $X = (X_1, X_2)$ is continuous with joint p.d.f. f_X. Recall from Theorem 1.307 that $f_X(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2), \forall x_1, x_2, \in \mathbb{R}$.

Then,

$$\begin{aligned}\mathbb{E}\left(\prod_{i=1}^{2} h_i(X_i)\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_1(x_1)h_2(x_2)f_X(x_1,x_2)\,dx_1dx_2\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_1(x_1)h_2(x_2)f_{X_1}(x_1)f_{X_2}(x_2)\,dx_1dx_2\\ &= \left(\int_{-\infty}^{\infty} h_1(x_1)f_{X_1}(x_1)\,dx_1\right)\left(\int_{-\infty}^{\infty} h_2(x_2)f_{X_2}(x_2)\,dx_2\right)\\ &= \prod_{i=1}^{2}\mathbb{E}h_i(X_i).\end{aligned}$$

(g) This is a special case of statement (f). Let $A_1, A_2, \cdots, A_p \subseteq \mathbb{R}$. Consider the functions

$$h_i(x_i) := \begin{cases} 1, & \text{if } x \in A_i \\ 0, & \text{otherwise.}\end{cases} = 1_{A_i}(x_i), \forall x_i \in \mathbb{R}, i = 1, 2, \cdots, p.$$

Note that $\mathbb{E}h_i(X_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} 1_{A_i}(x_i)f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_p}(x_p)\,dx_1dx_2\cdots dx_p = \int_{-\infty}^{\infty} 1_{A_i}(x_i)f_{X_i}(x_i)\,dx_i = \mathbb{P}(X_i \in A_i)$, when X is continuous. The same equality is also true when X is discrete. Now, consider the function $h : \mathbb{R}^p \to \mathbb{R}$ defined by $h(x) = \prod_{i=1}^{p} h_i(x_i), \forall x \in \mathbb{R}^p$. Using (f), we have

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \cdots, X_p \in A_p) = \prod_{i=1}^{p}\mathbb{P}(X_i \in A_i).$$

(h) Continue with the assumptions of statement (f). For fixed $y_1, y_2, \cdots, y_p \in \mathbb{R}$, consider the functions $g_1, g_2, \cdots, g_p : \mathbb{R} \to \mathbb{R}$ defined by

$$g_i(x_i) := \begin{cases} 1, & \text{if } h_i(x_i) \le y_i \\ 0, & \text{otherwise.}\end{cases} \quad \forall x_i \in \mathbb{R}, i = 1, 2, \cdots, p.$$

Note that $\mathbb{E}g_i(X_i) = \mathbb{P}(h_i(X_i) \le y_i) = F_{h_i(X_i)}(y_i), \forall i$ and

$$\begin{aligned}F_{h_1(X_1),h_2(X_2),\cdots,h_p(X_p)}(y_1, y_2, \cdots, y_p) &= \mathbb{P}(h_1(X_1) \le y_1, h_2(X_2) \le y_2, \cdots, h_p(X_p) \le y_p)\\ &= \prod_{i=1}^{p}\mathbb{P}(h_i(X_i) \le y_i)\end{aligned}$$

$$= \prod_{i=1}^{p} F_{h(X_i)}(y_i).$$

Hence, the RVs $h_1(X_1), h_2(X_2), \cdots, h_p(X_p)$ are independent.

(i) Let X_1, X_2 be independent RVs. Then $\mathbb{E}(X_1X_2) = (\mathbb{E}X_1)(\mathbb{E}X_2)$ and hence, using (d),

$$Cov(X_1, X_2) = 0.$$

Further, if $X_1, X_2, \cdots, X_p$ are independent, then using (e),

$$Var\left(\sum_{i=1}^{p} a_i X_i\right) = \sum_{i=1}^{p} a_i^2 Var\left(X_i\right)$$

for all real constants $a_1, a_2, \cdots, a_p$.

(j) Recall that $M_X : A \to \mathbb{R}$ is given by

$$M_X(t) = \mathbb{E}\left(e^{\sum_{i=1}^{p} t_i X_i}\right), \quad \forall t = (t_1, t_2, \ldots, t_p) \in A,$$

with

$$A := \left\{t = (t_1, t_2, \ldots, t_p) \in \mathbb{R}^p : \mathbb{E}\left(e^{\sum_{i=1}^{p} t_i X_i}\right) < \infty\right\}.$$

Taking $t = (0, 0, \cdots, 0) \in \mathbb{R}^p$ yields $M_X(0, 0, \cdots, 0) = 1$ and hence $(0, 0, \cdots, 0) \in A$. In particular, $A \neq \emptyset$. Also, $M_X(t) > 0, \forall t \in A$.

(k) If $t = (0, \cdots, 0, t_i, 0, \cdots, 0) \in A$, then $M_X(t) = \mathbb{E}\left(e^{\sum_{k=1}^{p} t_k X_k}\right) = \mathbb{E}\left(e^{t_i X_i}\right) = M_{X_i}(t_i)$. Similarly, if $t = (0, \cdots, 0, t_i, 0, \cdots, 0, t_j, 0, \cdots, 0) \in A$, then $M_X(t) = \mathbb{E}\left(e^{\sum_{k=1}^{p} t_k X_k}\right) = \mathbb{E}\left(e^{t_i X_i + t_j X_j}\right) = M_{X_i, X_j}(t_i, t_j)$.

(l) This result is being stated without proof. If $(-a_1, a_1) \times (-a_2, a_2) \times \cdots \times (-a_p, a_p) \subseteq A$ for some $a_1, a_2, \cdots, a_p > 0$, then M_X possesses partial derivatives of all orders in $(-a_1, a_1) \times (-a_2, a_2) \times \cdots \times (-a_p, a_p)$. Furthermore, for non-negative integers $k_1, \ldots, k_p$

$$\mathbb{E}\left(X_1^{k_1} X_2^{k_2} \cdots X_p^{k_p}\right) = \left[\frac{\partial^{k_1 + k_2 + k_3 + \cdots + k_p}}{\partial t_1^{k_1} \cdots \partial t_p^{k_p}} M_X(t)\right]_{(t_1, t_2 \ldots t_p) = (0, \ldots, 0)}.$$

For $i \neq j$ with $i, j \in \{1, \ldots, p\}$, we have

$$Cov\left(X_i, X_j\right)$$

$$
\begin{aligned}
&= \mathbb{E}\left(X_i X_j\right) - \mathbb{E}\left(X_i\right)\mathbb{E}\left(X_j\right) \\
&= \left[\frac{\partial^2}{\partial t_i \partial t_j} M_X(t)\right]_{(t_1,t_2\ldots t_p)=(0,\ldots,0)} - \left[\frac{\partial}{\partial t_i} M_X(t)\right]_{(t_1,t_2\ldots t_p)=(0,\ldots,0)} \left[\frac{\partial}{\partial t_j} M_X(t)\right]_{(t_1,t_2\ldots t_p)=(0,\ldots,0)} \\
&= \left[\frac{\partial^2}{\partial t_i \partial t_j} \Psi_X(t)\right],
\end{aligned}
$$

where $\Psi_X(t) := \ln M_X(t), t \in A$. Compare this with the one-dimensional case in Proposition 1.215.

(m) If $X_1, X_2, \cdots, X_p$ are independent, then for all $t \in A$,

$$
\begin{aligned}
M_X(t) &= \mathbb{E}\left(e^{\sum_{i=1}^p t_i X_i}\right) = \mathbb{E}\left(\prod_{i=1}^p e^{t_i X_i}\right) \\
&= \prod_{i=1}^p \mathbb{E}\left(e^{t_i X_i}\right) = \prod_{i=1}^p M_{X_i}(t_i).
\end{aligned}
$$

(n) If $(-a_1, a_1) \times (-a_2, a_2) \times \cdots \times (-a_p, a_p) \subseteq A$ for some $a_1, a_2, \cdots, a_p > 0$ and $M_X(t) = \prod_{i=1}^p M_{X_i}(t_i), \forall t \in A$, then it can be shown that $X_1, X_2, \cdots, X_p$ are independent. We do not discuss the proof of this result in this course.

Proposition 1.331 (Cauchy-Schwarz Inequality)**.** *Let X and Y be RVs defined on the same probability space. Then,*

$$(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2),$$

provided the expectations exist. The equality occurs if and only if $\mathbb{P}(Y = cX) = 1$ or $\mathbb{P}(X = cY) = 1$ for some $c \in \mathbb{R}$.

Proof. First we consider the case when $\mathbb{E}X^2 = 0$. Then $\mathbb{P}(X = 0) = 1$ and consequently $\mathbb{P}(XY = 0) = 1$ and $\mathbb{E}(XY) = 0$. The equality holds.

Now, assume that $\mathbb{E}X^2 > 0$. Now, for all $c \in \mathbb{R}$, we have $\mathbb{E}(Y - cX)^2 = c^2\mathbb{E}X^2 - 2c\mathbb{E}(XY) + \mathbb{E}Y^2 \geq 0$. Hence, the discriminant $(2\mathbb{E}(XY))^2 - 4\mathbb{E}X^2\mathbb{E}Y^2$ must be non-positive, which proves the statement.

If the equality holds for some $\mathbb{E}(Y - cX)^2 = 0$ for some c, then we have $\mathbb{P}(Y = cX) = 1$. If $\mathbb{P}(Y = cX) = 1$ for some c, then $\mathbb{E}(Y - cX)^2 = 0$. Interchanging the roles of X and Y, we can discuss the case involving $\mathbb{E}(X - cY)^2$ and $\mathbb{P}(X = cY)$. □

Corollary 1.332. *Let X and Y be RVs defined on the same probability space. Then,*

$$(Cov(X,Y))^2 \leq Var(X)\,Var(Y),$$

provided the covariance and the variances exist.

Proof. Take $U = X - \mathbb{E}X$ and $V = Y - \mathbb{E}Y$. Applying the Cauchy-Schwarz inequality to U and V, the result follows. □

Definition 1.333 (Correlation between RVs)**.** Let X and Y be RVs defined on the same probability space. If $0 < Var(X) < \infty, 0 < Var(Y) < \infty$, then we call

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

as the Correlation between X and Y. We say X and Y are uncorrelated if $\rho(X,Y) = 0$ or equivalently $Cov(X,Y) = 0$.

Note 1.334. By Corollary 1.332, $|\rho(X,Y)| \leq 1$ for any two RVs X and Y defined on the same probability space.

Remark 1.335 (Correlation and Independence). If X and Y are independent RVs defined on the same probability space, then by Remark 1.330(i), $Cov(X,Y) = 0$ and hence X and Y are uncorrelated. However, the converse is not true. We illustrate this problem with examples.

(a) Let $X = (X_1, X_2)$ be a bivariate discrete random vector, i.e. a 2-dimensional discrete random vector with joint p.m.f. given by

$$f_X(x_1,x_2) = \begin{cases} \frac{1}{2}, & \text{if } (x_1,x_2) = (0,0), \\ \frac{1}{4}, & \text{if } (x_1,x_2) = (1,1) \text{ or } (1,-1), \\ 0, & \text{otherwise.} \end{cases}$$

The marginal p.m.fs are

$$f_{X_1}(x_1) = \begin{cases} \frac{1}{2}, \text{ if } x_1 \in \{0,1\} \\ 0, \text{ otherwise} \end{cases} \quad , \quad f_{X_2}(x_2) = \begin{cases} \frac{1}{2}, \text{ if } x_2 = 0 \\ \frac{1}{4}, \text{ if } x_2 \in \{1,-1\} \\ 0, \text{ otherwise} \end{cases}$$

We have $f_{X_1,X_2}(0,0) = \frac{1}{2} \neq \frac{1}{4} = f_{X_1}(0)f_{X_2}(0)$ and hence X_1 and X_2 are not independent. But, $\mathbb{E}X_1 = \frac{1}{2}, \mathbb{E}X_2 = 0, \mathbb{E}(X_1X_2) = 0, Var(X_1) > 0$ and $Var(X_2) > 0$. Therefore $Cov(X_1, X_2) = 0$ and hence X_1 and X_2 are uncorrelated.

(b) Let $X = (X_1, X_2)$ be a bivariate continuous random vector, i.e. a 2-dimensional continuous random vector with joint p.d.f. given by

$$f_X(x_1, x_2) = \begin{cases} 1, & \text{if } 0 < |x_2| \leq x_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Then,

$$\mathbb{E}(X_1X_2) = \int_0^1 \int_{-x_1}^{x_1} x_1x_2 \, \mathrm{d}x_2 \, \mathrm{d}x_1 = 0,$$

and

$$\mathbb{E}(X_1) = \int_0^1 \int_{-x_1}^{x_1} x_1 \, \mathrm{d}x_2 \, \mathrm{d}x_1 = \frac{2}{3}, \quad \mathbb{E}(X_2) = \int_0^1 \int_{-x_1}^{x_1} x_2 \, \mathrm{d}x_2 \, \mathrm{d}x_1 = 0.$$

Hence, $\mathbb{E}(X_1X_2) = (\mathbb{E}X_1)(\mathbb{E}X_2)$, which implies $Cov(X_1, X_2) = 0$. A similar computation shows $Var(X_1)$ and $Var(X_2)$ exists and are non-zero. Hence, X_1 and X_2 are uncorrelated. Now, by computing the marginal p.d.f.s f_{X_1} and f_{X_2}, it is immediate that the equality

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

does not hold for all $x = (x_1, x_2) \in \mathbb{R}^2$. Here, X_1 and X_2 are not independent. The verification with the marginal p.d.f.s is left as an exercise in practice problem set 8.

We now discuss the concept of equality of distribution for random vectors. As we shall see, the ideas remain the same as in the case of RVs.

Definition 1.336 (Identically distributed random vectors)**.** Let X and Y be two p-dimensional random vectors, possibly defined on different probability spaces. We say that they have the same law/distribution, or equivalently, X and Y are identically distributed or equivalently, X and Y are equal in law/distribution, denoted by $X \overset{d}{=} Y$, if $F_X(x) = F_Y(x), \forall x \in \mathbb{R}^p$.

Remark 1.337. As discussed in the case of RVs in Remark 1.219, we can check whether two random vectors are identically distributed or not via other quantities that describe their law/distribution.

(a) Let X and Y be p-dimensional discrete random vectors with joint p.m.f.s f_X and f_Y, respectively. Then X and Y are identically distributed if and only if $f_X(x) = f_Y(x), \forall x \in \mathbb{R}^p$.

(b) Let X and Y be p-dimensional continuous random vectors with joint p.d.f.s f_X and f_Y, respectively. Then X and Y are identically distributed if and only if $f_X(x) = f_Y(x), \forall x \in \mathbb{R}^p$.

(c) Let X and Y be p-dimensional random vectors such that their joint MGFs M_X and M_Y exist and agree on $(-a_1, a_1) \times (-a_2, a_2) \times \cdots (-a_p, a_p)$ for some $a_1, a_2, \cdots, a_p > 0$, then X and Y are identically distributed.

(d) Let X and Y be identically distributed p-dimensional random vectors. Then for any function $h : \mathbb{R}^p \to \mathbb{R}^q$, we have $h(X)$ and $h(Y)$ are identically distributed q-dimensional random vectors.

Notation 1.338 (i.i.d RVs)**.** We say that RVs $X_1, \cdots, X_p$ defined on the same probability space are independent and identically distributed, then we usually use the short hand notation i.i.d. and say that $X_1, \cdots, X_p$ are i.i.d..

Note 1.339. We can generalize the concept of independence of RVs to independence of random vectors. To avoid notational complexity, we do not discuss this in this course.

Definition 1.340. (a) A random sample is a collection of i.i.d. RVs.

(b) A random sample of size n is a collection of n i.i.d. RVs $X_1, X_2, \cdots, X_n$.

(c) Let $X_1, X_2, \cdots, X_n$ be a random sample of size n. If the common DF is F or the common p.m.f./p.d.f. is f, then we call $X_1, X_2, \cdots, X_n$ to be a random sample from a distribution having a DF F or p.m.f./p.d.f. f.

(d) A function of one or more RVs that does not depend on any unknown parameter is called a statistic.

Example 1.341. Suppose that $X_1, \cdots, X_n$ are i.i.d. with the common distribution being $Poisson(\theta)$ or $Exponential(\theta)$ for some unknown $\theta \in (0, \infty)$. Here, θ is a unknown parameter.

(a) $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ is a statistic and is usually referred to as the sample mean.

(b) $X_1 - \theta$ is not a statistic.

(c) $S_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a statistic and is usually referred to as the sample variance. Depending on the situation, we sometimes work with $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

(d) The value of S_n such that S_n^2 is the sample variance, is referred to as the sample standard deviation.

(e) For $r = 1, \cdots, n$, we denote by $X_{(r:n)}$ the r-th smallest of $X_1, \cdots, X_n$. By definition, $X_{(1:n)} \leq \cdots \leq X_{(n:n)}$ and these are called the order statistics of the random sample. If n is understood, then we simply write $X_{(r)}$ to denote the r-th order statistic.

Note 1.342. Let X_1, X_2 be a random sample of size 2. Then $X_{(1)} = \min\{X_1, X_2\} = \frac{1}{2}(X_1 + X_2) - \frac{1}{2}|X_1 - X_2|$ and $X_{(2)} = \max\{X_1, X_2\} = \frac{1}{2}(X_1 + X_2) + \frac{1}{2}|X_1 - X_2|$ are RVs. Using similar arguments, it follows that the order statistics from any random sample of size n are RVs. The joint distribution of the order statistics is therefore of interest.

Note 1.343. Let $X_1, \cdots, X_n$ be a random sample of continuous RVs with the common p.d.f. f. Then,

$$\mathbb{P}(X_{(1)} < X_{(2)}) < \cdots < X_{(n)}) = 1$$

and hence $X_{(r)}, r = 1, \cdots, n$ are defined uniquely with probability one.

Proposition 1.344. *Let $X_1, \cdots, X_n$ be a random sample of continuous RVs with the common DF F and the common p.d.f. f. The joint p.d.f. of $(X_{(1)}, \cdots, X_{(n)})$ is given by*

$$g(y_1, \cdots, y_n) = \begin{cases} n! \prod_{i=1}^n f(y_i), & \textit{if } y_1 < \cdots < y_n, \\ 0, & \textit{otherwise.} \end{cases}$$

Further the marginal p.d.f. of $X_{(r)}$ is given by

$$g_{X_{(r)}}(y) = \frac{n!}{(r-1)!(n-r)!}(F(y))^{r-1}(1-F(y))^{n-r}f(y), \forall y \in \mathbb{R}.$$

Proof. Observe that a sample value $(y_1, \cdots, y_n)$ of $(X_{(1)}, \cdots, X_{(n)})$ is related to a sample $(x_1, \cdots, x_n)$ of $(X_1, \cdots, X_n)$ in the following way

$$(y_1, \cdots, y_n) = (x_{(1)}, \cdots, x_{(n)}),$$

$x_{(r)}$ being the r-th smallest of $x_1, \cdots, x_n$. Note that $y_r = x_{(r)}$.

Now, the actual values $x_1, \cdots, x_n$ may have been arranged in a different order than $x_{(1)}, \cdots, x_{(n)}$. In fact, the values $x_{(1)}, \cdots, x_{(n)}$ arise from one of the $n!$ permutations of the values $x_1, \cdots, x_n$. But, any such transformation/permutation is obtained by the action of a permutation matrix on the vector $(x_1, \cdots, x_n)$. For example, if $x_1 < x_2 < \cdots < x_{n-2} < x_n < x_{n-1}$, then $x_{(1)} = x_1, \cdots, x_{(n-2)} = x_{n-2}, x_{(n-1)} = x_n, x_{(n)} = x_{n-1}$ which interchanges the $n-1$ and n-th values, i.e. x_{n-1} and x_n.

Hence, the Jacobian matrix for this transformation is the same as the corresponding permutation matrix and the Jacobian determinant is ± 1.

Since $X_1, \cdots, X_n$ are i.i.d., the joint p.d.f. of $(X_1, \cdots, X_n)$ is given by

$$f_{X_1, \cdots, X_n}(x_1, \cdots, x_n) = f(x_1) \times \cdots \times f(x_n), \forall (x_1, \cdots, x_n) \in \mathbb{R}^n.$$

Using Theorem 1.322, we have joint p.d.f. of $(X_{(1)}, \cdots, X_{(n)})$ is given by

$$g(y_1, \cdots, y_n) = \begin{cases} n! \prod_{i=1}^n f(y_i), & \text{if } y_1 < \cdots < y_n, \\ 0, & \text{otherwise.} \end{cases}$$

The marginal p.d.f. of $X_{(r)}$ can now be computed for $y \in \mathbb{R}$,

$$g_{X_{(r)}}(y)$$

$$= \int_{y_{r-1}=-\infty}^{y} \int_{y_{r-2}=-\infty}^{y_{r-1}} \cdots \int_{y_1=-\infty}^{y_2} \int_{y_{r+1}=y}^{\infty} \int_{y_{r+2}=y_{r+1}}^{\infty} \cdots \int_{y_n=y_{n-1}}^{\infty} n! \prod_{i=1}^{n} f(y_i)\, dy_n dy_{n-1} \cdots dy_{r+1} dy_1 dy_2 \cdots dy_{r-1}$$

The above integral simplifies to the result stated above. □

Example 1.345. Let X_1, X_2, X_3 be a random sample from $Uniform(0,1)$ distribution. The common p.d.f. here is given by

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

By the above result, the joint p.d.f. of $(X_{(1)}, X_{(2)}, X_{(3)})$ is given by

$$g(y_1, y_2, y_3) = \begin{cases} 6, \text{ if } 0 < y_1 < y_2 < y_3 < 1, \\ 0, \text{ otherwise.} \end{cases}$$

and the marginal p.d.f. of $X_{(1)}$ is

$$g(y_1) = \begin{cases} 3(1-y_1)^2, \text{ if } y_1 \in (0,1) \\ 0, \text{ otherwise.} \end{cases}$$

Remark 1.346. For random samples from discrete distributions, there is no general formula or result which helps in computing the joint distribution of the order statistics. Usually they are done by a case-by-case analysis. Let X_1, X_2, X_3 be a random sample from $Bernoulli(p)$ distribution, for some $p \in (0,1)$. The common p.m.f. here is given by

$$f(x) = \begin{cases} p, \text{ if } x = 1 \\ 1-p, \text{ if } x = 0 \\ 0, \text{ otherwise.} \end{cases}$$

Note that $X_{(1)}$ is also a $\{0,1\}$-valued RV with $X_{(1)} = \min\{X_1, X_2, X_3\} = 1$ if and only if $X_1 = X_2 = X_3 = 1$. Then using independence,

$$\mathbb{P}(X_{(1)} = 1) = \mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 1) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1)\mathbb{P}(X_3 = 1) = p^3$$

and $\mathbb{P}(X_{(1)} = 0) = 1 - \mathbb{P}(X_{(1)} = 1) = 1 - p^3$. Therefore, $X_{(1)} \sim Bernoulli(p^3)$. Similarly, $X_{(3)} \sim Bernoulli(1 - (1-p)^3)$. The distribution of $X_{(2)}$ is left as an exercise in problem set 8.

Earlier, in Week 7, we have discussed the concept of conditional distributions. In this week, we have also discussed the concept of expectation of a random vector. Combining these two concepts, we are led to the following.

Definition 1.347 (Conditional Expectation, Conditional Variance and Conditional Covariance)**.** Let $X = (X_1, X_2, \cdots, X_{p+q})$ be a $p+q$-dimensional random vector with joint p.m.f./p.d.f. f_X. Let the joint p.m.f./p.d.f. $Y = (X_1, X_2, \cdots, X_p)$ and $Z = (X_{p+1}, X_{p+2}, \cdots, X_{p+q})$ be denoted by f_Y and f_Z, respectively. Let $h : \mathbb{R}^p \to \mathbb{R}$ be a function. Let $z \in \mathbb{R}^q$ be such that $f_Z(z) > 0$.

(a) The conditional expectation of $h(Y)$ given $Z = z$, denoted by $\mathbb{E}(h(Y) \mid Z = z)$, is the expectation of $h(Y)$ under the conditional distribution of Y given $Z = z$.

(b) The conditional variance of $h(Y)$ given $Z = z$, denoted by $Var(h(Y) \mid Z = z)$, is the variance of $h(Y)$ under the conditional distribution of Y given $Z = z$.

(c) Let $1 \leq i \neq j \leq p$. The conditional covariance between X_i and X_j given $Z = z$, denoted by $Cov(X_i, X_j \mid Z = z)$, is the covariance between X_i and X_j under the conditional distribution of (X_i, X_j) given $Z = z$.

Notation 1.348. On $\{z \in \mathbb{R}^q : f_Z(z) > 0\}$, consider the function, $g_1(z) := \mathbb{E}(h(Y) \mid Z = z)$. We denote the RV $g_1(Z)$ by $\mathbb{E}(h(Y) \mid Z)$. Similarly, define the RVs $Var(h(Y) \mid Z)$ and $Cov(X_1, X_2 \mid Z)$

Proposition 1.349. *The following are properties of Conditional Expectation, Conditional Variance and Conditional Covariance. Here, we assume that the relevant expectations exist.*

(a) $\mathbb{E}h(Y) = \mathbb{E}(\mathbb{E}(h(Y) \mid Z))$.

(b) $Var(h(Y)) = Var(\mathbb{E}(h(Y) \mid Z)) + \mathbb{E}Var(h(Y) \mid Z)$.

(c) $Cov(X_1, X_2) = Cov(\mathbb{E}(X_1 \mid Z), \mathbb{E}(X_2 \mid Z)) + \mathbb{E}Cov(X_1, X_2 \mid Z)$.

Proof. We only prove the first statement under a simple assumption. The general case and other statements can be proved using appropriate generalization.

Take $p = q = 1$ and let $X = (Y, Z)$ be a 2-dimensional continuous random vector. Then,

$$\begin{aligned}\mathbb{E}(\mathbb{E}(h(Y) \mid Z) &= \int_{-\infty}^{\infty} \mathbb{E}(h(Y) \mid Z = z) f_Z(z)\, dz \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} h(y) f_{Y|Z}(y \mid z)\, dy\right] f_Z(z)\, dz \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(y) f_{Y,Z}(y, z)\, dydz \\ &= \mathbb{E}h(Y).\end{aligned}$$

□

Example 1.350. We shall see computations for conditional expectations in a later lecture.

We look at examples of discrete RVs in relation with random experiments.

Remark 1.351 (Binomial RVs via random experiments). Recall that in Remark 1.228, we have seen Bernoulli RVs arising from Bernoulli trials. Now, consider the same random experiment with two outcomes 'Success' and 'Failure' with probability of success $p \in (0, 1)$. Now, consider n independent Bernoulli trials of this experiment with the RV X_i being 1 for 'Success' and 0 for 'Failure' in the i-th trial for $i = 1, 2, \cdots, n$. Then, $X_1, X_2, \cdots, X_n$ is a random sample of size n from the $Bernoulli(p)$ distribution. Now, the total number X of successes in the n trials is given by $X = X_1 + X_2 + \cdots + X_n$ and hence, by Remark 1.321, $X \sim Binomial(n, p)$. A $Binomial(n, p)$ RV can therefore be interpretated as the number of successes in n trials of a random experiment with two outcomes 'Success' and 'Failure' with probability of success $p \in (0, 1)$. Here, we have kept p fixed over all the trials.

Example 1.352. Suppose that a standard six-sided fair die is rolled at random 4 times independently. We now consider the probability that all the rolls result in a number at least 5. In each roll, obtaining at least 5 has the probability $\frac{2}{6} = \frac{1}{3}$ - we treat this as the probability of success in one trial. Repeating the trial three times independently gives us the number of success as

$X \sim \mathit{Binomial}(4, \frac{1}{3})$. The probability that all the rolls result in successes is given by $\mathbb{P}(X = 4)$ – which can now be computed from the Binomial distribution. If we now consider the probability that at least two rolls result in a number at least 5, then that probability is given by $\mathbb{P}(X \geq 2)$.

Example 1.353 (Negative Binomial RV)**.** Consider a random experiment with two outcomes 'Success' and 'Failure' with probability of success $p \in (0, 1)$. We consider repeating the experiment until we have r successes, with r being a positive integer. Let X denote the number of failures observed till the r-th success. Then X is a discrete RV with the support of X being $S_X = \{0, 1, \cdots\}$. Note that for $k \in S_X$, using independence of the trials we have

$$\begin{aligned}
&\mathbb{P}(X = k) \\
&= \mathbb{P}(\text{there are } k \text{ failures before the } r\text{-th success}) \\
&= \mathbb{P}(\text{first } k + r - 1 \text{ trials result in } r - 1 \text{ successes and the } k + r\text{-th trial results in a success}) \\
&= \mathbb{P}(\text{first } k + r - 1 \text{ trials result in } r - 1 \text{ successes}) \times \mathbb{P}(\text{the } k + r\text{-th trial results in a success}) \\
&= \binom{k + r - 1}{r - 1} p^{r-1}(1 - p)^k \times p \\
&= \binom{k + r - 1}{k} p^r (1 - p)^k.
\end{aligned}$$

Therefore the p.m.f. of X is given by

$$f_X(x) = \begin{cases} \binom{x+r-1}{x} p^r (1 - p)^x, & \text{if } x \in S_X, \\ 0, & \text{otherwise.} \end{cases}$$

In this case, we say X follows the negative Binomial(r, p) distribution or equivalently, X is a negative Binomial (r, p) RV. Here, r denotes the number of successes at which the trials are terminated and p being the probability of success. The MGF can now be computed as follows.

$$\begin{aligned}
M_X(t) &= \mathbb{E}e^{tX} \\
&= \sum_{k=0}^{\infty} e^{tk} \binom{k + r - 1}{k} p^r (1 - p)^k
\end{aligned}$$

$$= \sum_{k=0}^{\infty} \binom{k+r-1}{k} p^r \left[(1-p)e^t\right]^k$$
$$= p^r[1-(1-p)e^t]^{-r}, \forall t < -\ln(1-p).$$

Using the MGF, we can compute the mean and variance of X as $\mathbb{E}X = \frac{rq}{p}, Var(X) = \frac{rq}{p^2}$, with $q = 1-p$.

Remark 1.354 (Connection between negative Binomial distribution and the Geometric distribution). A negative Binomial$(1, p)$ RV X has the p.m.f.

$$f_X(x) = \begin{cases} p(1-p)^x, \text{ if } x \in \{0, 1, \cdots\}, \\ 0, \text{ otherwise.} \end{cases}$$

which is exactly the same as the p.m.f. for the Geometric(p) distribution. Since the p.m.f. of a discrete RV determines the distribution, we conclude that a *Geometric*(p) RV can be identified as the number of failures observed till the first success in independent trials of a random experiment with two outcomes 'Success' and 'Failure' with probability of success $p \in (0,1)$.

Note 1.355 (No memory property for Geometric Distribution)**.** Let $X \sim$ *Geometric*(p) for some $p \in (0,1)$. For any non-negative integer n, we have

$$\mathbb{P}(X \geq n) = \sum_{k=n}^{\infty} p(1-p)^k = p(1-p)^n \sum_{k=0}^{\infty} (1-p)^k = (1-p)^n.$$

Then, for any non-negative integers m, n, we have

$$\mathbb{P}(X \geq m+n \mid X \geq m) = \frac{\mathbb{P}(X \geq m+n \text{ and } X \geq m)}{\mathbb{P}(X \geq m)} = \frac{\mathbb{P}(X \geq m+n)}{\mathbb{P}(X \geq m)} = (1-p)^n = \mathbb{P}(X \geq n).$$

Here, the probability of obtaining at least n additional failures (till the first success) beyond the first m or more failures remain the same as in the the probability of obtaining at least n failures till the first success. In the situation where we stress test a device under repeated shocks, if we consider the survival or continued operation of the device under shocks as 'Failures' in our trial and if the number of shocks till the device breaks down follows *Geometric*(p) distribution, then we can interpret that the age of the device (measured in number of shocks observed) has no effect on the remaining lifetime of the device. This property is usually referred to as the 'No memory' property of the Geometric distribution.

Note 1.356. See problem set 8 for a similar property for the Exponential distribution.

Example 1.357. Let us consider the random experiment of rolling a standard six-sided fair die till we observe an outcome of at least 5. As mentioned in Example 1.352, the probability of success is $\frac{1}{3}$. Since the last roll results in a success, the number Y of rolls required is exactly one more than the number X of failures observed. Here $X \sim \mathit{Geometric}(\frac{1}{3})$. Then, the probability that an outcome of 5 or 6 is observed in the 10-th roll for the first time is given by

$$\mathbb{P}(Y = 10) = \mathbb{P}(X = 9) = \frac{1}{3}\left(\frac{2}{3}\right)^9.$$

If we want to look at Z which is the number of failures observed till 5 or 6 is rolled twice, then Z follows negative Binomial$(2, \frac{1}{3})$. Now, the number of rolls required is $Z + 2$. The probability that 10 rolls are required is given by

$$\mathbb{P}(Z + 2 = 10) = \mathbb{P}(Z = 8) = \binom{9}{8}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^8.$$

Note 1.358. Suppose that a box contains N items, out of which M items have been marked/labelled. In our experiment, we consider all labelled items to be identical and the same for all the unlabelled items. If we draw items from the box with replacement, then the probability of drawing a marked/labelled item is $\frac{M}{N}$ does not change between the draws. If we draw n items at random with replacement, then the number X of marked/labelled items follow $\mathit{Binomial}(n, \frac{M}{N})$ distribution. The case where the draws are conducted without replacement is of interest.

Example 1.359 (Hypergeometric RV)**.** In the setup of Note 1.358, consider drawing n items at random without replacement. Here, the probability of drawing a marked/labelled item may change between the draws and the number X of marked/labelled items in the n drawn items need not follow $\mathit{Binomial}(n, \frac{M}{N})$ distribution. Here, the number of labelled items among the items drawn satisfies the relation

$$0 \leq X \leq \min\{n, M\} \leq N$$

and the number of unlabelled items among the items drawn satisfies the relation

$$0 \leq n - X \leq N - M$$

and hence X is a discrete RV with support $S_X = \{\max\{0, n-(N-M)\}, \max\{0, n-(N-M)\}+1, \cdots, \min\{n, M\}\}$. The p.m.f. of X is given by

$$f_X(x) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, & \text{if } x \in S_X, \\ 0, & \text{otherwise.} \end{cases}$$

In this case, we say X follows the Hypergeometric distribution or equivalently, X is a Hypergeometric RV. This distribution has the three parameters N, M and n. Using properties of binomial coefficients, we can compute the factorial moments of X (left as exercise in problem set 9) and using these values we have,

$$\mathbb{E}X = \frac{nM}{N}, \quad Var(X) = \frac{nM}{N^2(N-1)}(N-M)(N-n) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}.$$

Note 1.360. In the setup of a Hypergeometric RV, if we consider $p = \frac{M}{N}$ as the probability of success and n as the number of trials, then $\mathbb{E}X$ matches with that of a $Binomial(n, \frac{M}{N})$ RV and $Var(X)$ is close to that of a $Binomial(n, \frac{M}{N})$ RV for small sample sizes n.

Example 1.361. Suppose that there are multiple boxes each containing 100 electric bulbs and we draw 5 bulbs from each box for testing. If a box contains 10 defective bulbs, then the number X of defective bulbs in the drawn bulbs follows Hypergeometric distribution with parameters $N = 100, M = 10, n = 5$. Here,

$$\mathbb{P}(X = 2) = \frac{\binom{10}{2}\binom{100-10}{5-2}}{\binom{100}{5}}.$$

Note 1.362. We continue with the setting of Note 1.358, where a box contains N items, out of which M items have been marked/labelled or are defective. In our experiment, we consider all labelled items to be identical and the same for all the unlabelled items. If we draw items from the box with replacement until the r-th defective item is drawn, then the number of draws required can be described in terms of negative Binomial$(r, \frac{M}{N})$ distribution, where the last draw yields the r-th defective item (see Example 1.357). The case where the draws are conducted without replacement is of interest.

Example 1.363 (Negative Hypergeometric RV)**.** In the setting of Note 1.362, consider drawing the items without replacement till the r-th defective item is obtained. We then have $1 \leq r \leq M$. Let X be the number of draws required. Then X is a discrete RV with support $S_X = \{r, r+1, \cdots, N\}$. For $k \in S_X$, using independence of the draws we have

$$
\begin{aligned}
&\mathbb{P}(X = k) \\
&= \mathbb{P}(\text{first } k-1 \text{ trials result in } r-1 \text{ defective items and the } k\text{-th trial results in a defective item}) \\
&= \mathbb{P}(\text{first } k-1 \text{ trials result in } r-1 \text{ defective items}) \times \mathbb{P}(\text{the } k\text{-th trial results in a defective item}) \\
&= \frac{\binom{M}{r-1}\binom{N-M}{k-r}}{\binom{N}{k-1}} \times \frac{M-(r-1)}{N-(k-1)}.
\end{aligned}
$$

Therefore the p.m.f. of X is given by

$$
f_X(x) = \begin{cases} \frac{M-(r-1)}{N-(x-1)} \frac{\binom{M}{r-1}\binom{N-M}{x-r}}{\binom{N}{x-1}}, & \text{if } x \in \{r, r+1, \cdots, N\}, \\ 0, & \text{otherwise.} \end{cases}
$$

In this case, we say X follows the negative Hypergeometric distribution or equivalently, X is a negative Hypergeometric RV.

We now discuss an example of a discrete random vector.

Remark 1.364. While considering a Bernoulli or a Binomial RV, we looked at random experiments with exactly two outcomes. We now consider random experiments with two or more than two outcomes. Suppose a random experiment terminates in one of k possible outcomes $A_1, A_2, \cdots, A_k$ for $k \geq 2$. More generally, we may also consider random experiments which terminate in one of k mutually exclusive and exhaustive events $A_1, A_2, \cdots, A_k$ with $k \geq 2$. Write $p_j = \mathbb{P}(A_j), j = 1, 2, \cdots, k$, which does not change from trial to trial. Then, $p_1 + p_2 + \cdots + p_k = 1$. Suppose n trials are conducted independently and let $X_j, j = 1, 2, \cdots, k$ denote the number of times event A_j has occured, respectively. Then the RVs $X_1, X_2, \cdots, X_k$ satisfy the relation $X_1 + X_2 + \cdots + X_k = n$ and we have

$$\mathbb{P}(X_1 = x_1, \cdots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

for $x_1, \cdots, x_k \in \{0, 1, \cdots, n\}$ with $x_1 + \cdots + x_k = n$. The probability is zero otherwise. Removing the redundancy we have the joint p.m.f. of $(X_1, X_2, \cdots, X_{k-1})$ given by

$$\begin{aligned} &f_{X_1, \cdots, X_{k-1}}(x_1, \cdots, x_{k-1}) \\ &= \frac{n!}{x_1! \cdots x_{k-1}!(n - x_1 - \cdots - x_{k-1})!} p_1^{x_1} \cdots p_{k-1}^{x_{k-1}} (1 - p_1 - \cdots - p_{k-1})^{n - x_1 - \cdots - x_{k-1}} \end{aligned}$$

for $x_1, \cdots, x_k \in \{0, 1, \cdots, n\}$ with $x_1 + \cdots + x_{k-1} \leq n$ and zero otherwise.

Example 1.365 (Multinomial Distribution)**.** A random vector $X = (X_1, \cdots, X_{k-1})$ is said to follow the Multinomial distribution with parameters n and $p_1, p_2, \cdots, p_k$ if the joint p.m.f. is as in Remark 1.364 above. We now list some properties of the multinomial distribution.

(a) We first compute the joint MGF. For $t_1, t_2, \cdots, t_{k-1} \in \mathbb{R}$,

$$\begin{aligned} &M_X(t_1, t_2, \cdots, t_{k-1}) \\ &= \mathbb{E} \exp(t_1 X_1 + t_2 X_2 + \cdots + t_{k-1} X_{k-1}) \\ &= \sum_{\substack{x_1, \cdots, x_k \in \{0, 1, \cdots, n\} \\ x_1 + \cdots + x_{k-1} \leq n}} \frac{n! \exp(t_1 x_1 + t_2 x_2 + \cdots + t_{k-1} x_{k-1})}{x_1! \cdots x_{k-1}!(n - x_1 - \cdots - x_{k-1})!} p_1^{x_1} \cdots p_{k-1}^{x_{k-1}} p_k^{n - x_1 - \cdots - x_{k-1}} \\ &= \sum_{\substack{x_1, \cdots, x_k \in \{0, 1, \cdots, n\} \\ x_1 + \cdots + x_{k-1} \leq n}} \frac{n!}{x_1! \cdots x_{k-1}!(n - x_1 - \cdots - x_{k-1})!} \left(p_1 e^{t_1}\right)^{x_1} \cdots \left(p_{k-1} e^{t_{k-1}}\right)^{x_{k-1}} p_k^{n - x_1 - \cdots - x_{k-1}} \\ &= \left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_{k-1} e^{t_{k-1}} + p_k\right)^n \end{aligned}$$

(b) If $t = (t_1, 0, \cdots, 0) \in \mathbb{R}^{k-1}$, then $M_X(t) = \mathbb{E} \exp(t_1 X_1) = M_{X_1}(t_1)$. But, using the above expression for the joint MGF, we have $M_{X_1}(t_1) = M_X(t) = (p_1 e^{t_1} + p_2 + \cdots + p_{k-1} + p_k)^n = (p_1 e^{t_1} + 1 - p_1)^n$. Therefore, $X_1 \sim Binomial(n, p_1)$. Similarly, $X_i \sim Binomial(n, p_i), \forall i = 2, \cdots, k-1$. In particular, $\mathbb{E} X_i = np_i, Var(X_i) = np_i(1 - p_i)$.

(c) For distinct indices $i, j \in \{1, 2, \cdots, k-1\}$,

$$M_{X_i,X_j}(t_i,t_j) = M_X(0,\cdots,0,t_i,0,\cdots,0,t_j,0,\cdots,0) = \left(p_i e^{t_i} + p_j e^{t_j} + 1 - p_i - p_j\right)^n, \forall (t_i,t_j) \in \mathbb{R}^2.$$

Therefore (X_i, X_j) follows the trinomial distribution with the parameters $p_i, p_j, 1-p_i-p_j$, i.e. multinomial distribution with the parameters $n=3$ and $p_i, p_j, 1-p_i-p_j$.

(d) For distinct indices $i, j \in \{1, 2, \cdots, k-1\}$, consider $t_i = t_j = t \in \mathbb{R}$. Then,

$$M_{X_i+X_j}(t) = M_{X_i,X_j}(t,t) = \left[(p_i+p_j)\, e^t + 1 - (p_i+p_j)\right]^n,$$

which shows $X_i + X_j \sim Binomial(n, p_i + p_j)$. Then $Var(X_i+X_j) = n(p_i+p_j)(1-p_i-p_j)$. Using the relation

$$Var(X_i+X_j) = Var(X_i) + Var(X_j) + 2Cov(X_i,X_j),$$

we have $Cov(X_i, X_j) = -np_ip_j$. Consequently, the correlation between X_i and X_j is

$$\rho(X_i,X_j) = \frac{Cov(X_i,X_j)}{\sqrt{Var(X_i)\,Var(X_j)}} = -\left(\frac{p_ip_j}{(1-p_i)(1-p_j)}\right)^{\frac{1}{2}}.$$

Note 1.366. We now look at distributions that arise in practice from random samples. Such distributions are usually referred to as sampling distributions. More specifically, if $X_1, X_2, \cdots, X_n$ is a random sample from $N(\mu, \sigma^2)$ distribution, we shall look at various statistics involving the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

Note 1.367 (Distribution of square of a standard Normal RV)**.** Let $X \sim N(0,1)$. Recall that the p.d.f. of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \forall x \in \mathbb{R}$$

We consider the distribution of $Y = X^2$ by first computing the MGF. We have,

$$M_Y(t) = \mathbb{E}e^{tX^2} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tx^2} e^{\left(-\frac{x^2}{2}\right)}\, dx = \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^{\infty} e^{\left(t-\frac{1}{2}\right)x^2}\, dx = (1-2t)^{-\frac{1}{2}}, \forall t < \frac{1}{2}$$

Comparing with the MGF of the $Gamma(\alpha, \beta)$ distribution, we conclude that $X^2 \sim Gamma(\frac{1}{2}, 2)$.

Note 1.368. If $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$ and hence $\left(\frac{X-\mu}{\sigma}\right)^2 \sim Gamma(\frac{1}{2}, 2)$.

Remark 1.369 (Distribution of the sample mean for a random sample from the Normal distribution)*.* If $X_1, X_2, \cdots, X_n$ is a random sample from $N(\mu, \sigma^2)$ distribution, then for $Y = X_1 + X_2 + \cdots + X_n$, using independence of X_i's we have

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \exp(n\mu t + \frac{1}{2}n\sigma^2 t^2)$$

and hence $X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$. Now, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and consequently, $\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \sim N(0,1)$ and $n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 \sim Gamma(\frac{1}{2}, 2)$.

Note 1.370. Let $X_1, X_2, \cdots, X_n$ be independent RVs with $X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2, \cdots, n$. Then $\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2, i = 1, 2, \cdots, n$ are i.i.d. with the common distribution $Gamma(\frac{1}{2}, 2)$. The independence follows from Remark 1.330(h). Using problem set 8, we have

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

Definition 1.371 (Chi-Squared distribution with n degrees of freedom)**.** Let n be a positive integer. We refer to the $Gamma\left(\frac{n}{2}, 2\right)$ distribution as the Chi-Squared distribution with n degrees of freedom. If an RV X follows the Chi-Squared distribution with n degrees of freedom, we write $X \sim \chi_n^2$.

Note 1.372. Using Note 1.370, we conclude that $\sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2 \sim \chi_n^2$, where $X_1, X_2, \cdots, X_n$ are independent RVs with $X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2, \cdots, n$.

Note 1.373. As argued in Note 1.370, using Remark 1.330(h) we conclude that $X + Y \sim \chi_{m+n}^2$, where X, Y are independent RVs with $X \sim \chi_m^2$ and $Y \sim \chi_n^2$.

Note 1.374. If $X \sim \chi_n^2$, then using properties of the $Gamma\left(\frac{n}{2}, 2\right)$ distribution, we have $\mathbb{E}X = n, Var(X) = 2n$ and $M_X(t) = (1-2t)^{-\frac{n}{2}}, \forall t < \frac{1}{2}$.

Remark 1.375 (Distribution of the sample variance for a random sample from the Normal distribution)*.* Let $X_1, X_2, \cdots, X_n$ be a random sample from $N(\mu, \sigma^2)$ distribution. By looking at the joint MGF of $X_1 - \bar{X}, \cdots, X_n - \bar{X}$ and $\bar{X}$, it can be shown that $\sum_{i=1}^{n}(X_i - \bar{X})^2$ and $\bar{X}$ are independent.

Now,

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\bar{X}+\bar{X}-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\bar{X})^2 + \frac{n(\bar{X}-\mu)^2}{\sigma^2}$$

where $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)^2 \sim \chi_n^2$ and $\frac{n(\bar{X}-\mu)^2}{\sigma^2} \sim \chi_1^2$. Since, $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)^2$ and $\frac{n(\bar{X}-\mu)^2}{\sigma^2}$ are independent, we conclude $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\bar{X})^2 \sim \chi_{n-1}^2$. Taking the sample variance as $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2$, we conclude that $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$.

Note 1.376. Given a random sample $X_1, X_2, \cdots, X_n$ from $N(\mu, \sigma^2)$ distribution, the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2$ has the property that $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$. The distribution of $\frac{\bar{X}-\mu}{S_n}$ is of interest.

Definition 1.377 (Student's t-distribution with n degrees of freedom)**.** Let n be a positive integer. Let $X \sim N(0,1)$ and $Y \sim \chi_n^2$ be independent RVs. Then,

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

is said to follow the t-distribution with n degrees of freedom. In this case, we write $T \sim t_n$. The p.d.f. is given by

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)\sqrt{n}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \forall t \in \mathbb{R}.$$

Here, $\mathbb{E}T^k$ exists if $k < n$. Since, the distribution is symmetric about 0 and hence $\mathbb{E}T^k = 0$ for all k odd with $k < n$. If k is even and $k < n$, then

$$\mathbb{E}T^k = n^{\frac{k}{2}}\frac{\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{n-k}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)}.$$

In particular, if $n > 2$, then $\mathbb{E}T = 0$ and $Var(T) = \frac{n}{n-2}$. The t-distribution appears in the tests for statistical significance.

Note 1.378. If $X_1, X_2, \cdots, X_n$ is a random sample from $N(\mu, \sigma^2)$ distribution, then $\sqrt{n}\frac{\bar{X}-\mu}{S_n} \sim t_{n-1}$.

Note 1.379. Let $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ be independent random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distribution, respectively. Consider the sample variances $S_1^2 := \frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X})^2$ and $S_2^2 := \frac{1}{n-1}\sum_{j=1}^n (Y_j - \bar{Y})^2$. The distribution of $\frac{S_1^2}{S_2^2}$ is of interest. Note that $\frac{(m-1)S_1^2}{\sigma_1^2} \sim \chi^2_{m-1}$ and $\frac{(n-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n-1}$.

Definition 1.380 (F-distribution with degrees of freedom m and n)**.** Let m and n be positive integers. Let $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ be independent RVs. Then,

$$F = \frac{\frac{X}{m}}{\frac{Y}{n}}$$

is said to follow the F-distribution with degrees of freedom m and n. In this case, we write $F \sim F_{m,n}$. The p.d.f. is given by

$$\begin{aligned} f_F(x) &= \begin{cases} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{m}{n} \left(\frac{m}{n}x\right)^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases} \\ &= \begin{cases} \frac{1}{B\left(\frac{m}{2},\frac{n}{2}\right)} \frac{m}{n} \left(\frac{m}{n}x\right)^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Note 1.381. If $F \sim F_{m,n}$, then $\frac{1}{F} \sim F_{n,m}$.

Note 1.382. Let $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ be independent random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distribution, respectively. Consider the sample variances $S_1^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X})^2$ and $S_2^2 = \frac{1}{n-1}\sum_{j=1}^n (Y_j - \bar{Y})^2$. The distribution of $\frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \sim F_{m-1,n-1}$.

We now discuss an example of a bivariate continuous random vector. Appropriate generalizations to higher dimensions are possible, but we do not discuss that here.

Definition 1.383 (Bivariate Normal distribution)**.** A bivariate random vector $X = (X_1, X_2)$ is said to follow bivariate Normal distribution $N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ for $\mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}, \sigma_1 > 0, \sigma_2 > 0, \rho \in (-1, 1)$, if the joint p.d.f. is given by

$$f_{X_1,X_2}(x_1, x_2)$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right],$$

for all $(x_1, x_2) \in \mathbb{R}^2$. To check that f_{X_1,X_2}, as above, is a p.d.f., first note that $f_{X_1,X_2}(x_1, x_2) \geq 0, \forall (x_1, x_2) \in \mathbb{R}^2$. Now, changing variables to $y_1 = \frac{x_1-\mu_1}{\sigma_1}, y_2 = \frac{x_2-\mu_2}{\sigma_2}$, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\, dx_1 dx_2 \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left(-\frac{y_2^2}{2}\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left[-\frac{1}{2(1-\rho^2)}(y_1-\rho y_2)^2\right] dy_1 dy_2 \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left(-\frac{y_2^2}{2}\right) dy_2 \\
&= 1.
\end{aligned}$$

This completes the verification that f_{X_1,X_2} is a joint p.d.f..

Remark 1.384. Let $X = (X_1, X_2) \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ for some $\mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}, \sigma_1 > 0, \sigma_2 > 0, \rho \in (-1, 1)$.

(a) The marginal p.d.f. of X_2 is given by

$$\begin{aligned}
&f_{X_2}(x_2) \\
&= \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\, dx_1, \forall x_2 \in \mathbb{R} \\
&= \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right] \\
&\times \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2\sigma_1^2(1-\rho^2)}\left\{x_1 - \left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\right)\right\}^2\right] dx_1 \\
&= \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]
\end{aligned}$$

and hence $X_2 \sim N(\mu_2, \sigma_2^2)$. Similarly, $X_1 \sim N(\mu_1, \sigma_1^2)$. Thus the parameters $\mu_1 = \mathbb{E}X_1, \sigma_1^2 = Var(X_1), \mu_2 = \mathbb{E}X_2, \sigma_2^2 = Var(X_2)$ have their own interpretation.

(b) The covariance $Cov(X_1, X_2)$ is given by

$$\begin{aligned} Cov(X_1, X_2) &= \mathbb{E}(X_1 - \mu_1)(X_2 - \mu_2) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \mu_1)(x_2 - \mu_2) f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 \end{aligned}$$

By changing variables to $y_1 = \frac{x_1-\mu_1}{\sigma_1}, y_2 = \frac{x_2-\mu_2}{\sigma_2}$, and simplifying the above expression, we have $Cov(X_1, X_2) = \rho\sigma_1\sigma_2$. Consequently, the correlation $\rho(X_1, X_2) = \rho$. We now have the interpretation of the parameter ρ.

(c) The conditional distribution of X_1 given $X_2 = x_2 \in \mathbb{R}$ is described by the conditional p.d.f.

$$\begin{aligned} f_{X_1|X_2}(x_1 \mid x_2) &= \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sigma_1^2(1-\rho^2)} \left\{x_1 - \left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)\right)\right\}^2\right], \forall x_1 \in \mathbb{R} \end{aligned}$$

and hence $X_1 \mid X_2 = x_2 \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2))$. Similarly, for $x_1 \in \mathbb{R}$, $X_2 \mid X_1 = x_1 \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2))$.

(d) Using the conditional distributions obtained above, we conclude

$$\begin{aligned} \mathbb{E}[X_1 \mid X_2 = x_2] &= \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \\ Var[X_1 \mid X_2 = x_2] &= \sigma_1^2(1-\rho^2), \\ \mathbb{E}[X_2 \mid X_1 = x_1] &= \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \\ Var[X_2 \mid X_1 = x_1] &= \sigma_2^2(1-\rho^2). \end{aligned}$$

(e) If X_1 and X_2 are independent, then they are uncorrelated and in particular $\rho = \rho(X_1, X_2) = 0$. Conversely, if $\rho = 0$, then

$$\begin{aligned} f_{X_1,X_2}(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{1}{2}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right] \\ &= f_{X_1}(x_1) f_{X_2}(x_2), \forall (x_1, x_2) \in \mathbb{R}^2 \end{aligned}$$

and hence X_1 and X_2 are independent.

(f) Consider the matrix

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

This is usually called the Covariance matrix of $X = (X_1, X_2)$. Alternative terminology such as Variance-Covariance matrix or Dispersion matrix is also used. Observe that this is a symmetric matrix with $det(\Sigma) = \sigma_1^2\sigma_2^2(1-\rho^2) > 0$ and hence the matrix is invertible. In fact, this matrix is positive-definite and its eigen-values are positive. The joint p.d.f. can be rewritten as

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2}(x_1 - \mu_1, x_2 - \mu_2)\Sigma^{-1}\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}\right], \forall (x_1, x_2) \in \mathbb{R}^2.$$

(g) We now compute the joint MGF of X. We have

$$\begin{aligned}
&M_X(t_1, t_2) \\
&= \mathbb{E}\exp(t_1X_1 + t_2X_2) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(t_1x_1 + t_2x_2) f_{X_1,X_2}(x_1, x_2)\, dx_1dx_2 \\
&= \int_{-\infty}^{\infty} \exp(t_2x_2) f_{X_2}(x_2) \int_{-\infty}^{\infty} \exp(t_1x_1) f_{X_1|X_2}(x_1 \mid x_2)\, dx_1dx_2 \\
&= \int_{-\infty}^{\infty} \exp(t_2x_2) f_{X_2}(x_2) \exp(\mu_1t_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)t_1 + \frac{1}{2}\sigma_1^2(1-\rho^2)t_1^2)\, dx_2 \\
&= \exp(\mu_1t_1 + \rho\frac{\sigma_1}{\sigma_2}(-\mu_2)t_1 + \frac{1}{2}\sigma_1^2(1-\rho^2)t_1^2) \int_{-\infty}^{\infty} \exp(t_2x_2 + \rho\frac{\sigma_1}{\sigma_2}t_1x_2) f_{X_2}(x_2)\, dx_2 \\
&= \exp(\mu_1t_1 + \rho\frac{\sigma_1}{\sigma_2}(-\mu_2)t_1 + \frac{1}{2}\sigma_1^2(1-\rho^2)t_1^2) \exp\left(\mu_2\left\{t_2 + \rho\frac{\sigma_1}{\sigma_2}t_1\right\} + \frac{1}{2}\sigma_2^2\left\{t_2 + \rho\frac{\sigma_1}{\sigma_2}t_1\right\}^2\right) \\
&= \exp(\mu_1t_1 + \mu_2t_2 + \frac{1}{2}\sigma_1^2t_1^2 + \frac{1}{2}\sigma_2^2t_2^2 + \rho\sigma_1\sigma_2t_1t_2), \forall (t_1, t_2) \in \mathbb{R}^2.
\end{aligned}$$

(h) Let $c_1, c_2 \in \mathbb{R}$ such that at least one of c_1, c_2 is not zero and take $Y = c_1X_1 + c_2X_2$. Now,

$$M_Y(t) = \mathbb{E}\exp(c_1tX_1 + c_2tX_2) = \exp\left[(\mu_1c_1 + \mu_2c_2)t + \left(\frac{1}{2}\sigma_1^2c_1^2 + \frac{1}{2}\sigma_2^2c_2^2 + \rho\sigma_1\sigma_2c_1c_2\right)t^2\right], \forall t \in \mathbb{R}.$$

Looking at the structure of the MGF, we conclude that $Y = c_1X_1 + c_2X_2 \sim N(c_1\mu_1 + c_2\mu_2, c_1^2\sigma_1^2 + c_2\sigma_2^2 + 2\rho c_1c_2\sigma_1\sigma_2)$.

Remark 1.385. The above statement for the linear combination of X_1, X_2 actually characterizes the bivariate Normal distribution. If $X = (X_1, X_2)$ is such that $\mathbb{E}X_1 = \mu_1, \mathbb{E}X_2 = \mu_2, Var(X_1) = \sigma_1^2 > 0, Var(X_2) = \sigma_2^2 > 0, \rho(X_1, X_2) = \rho \in (-1, 1)$ and $c_1X_1 + c_2X_2 \sim N(c_1\mu_1 + c_2\mu_2, c_1^2\sigma_1^2 + c_2\sigma_2^2 + 2\rho c_1c_2\sigma_1\sigma_2)$ for all $(c_1, c_2) \neq (0, 0)$, then

$$\begin{aligned} M_X(t_1, t_2) &= \mathbb{E}\exp(t_1X_1 + t_2X_2) \\ &= M_{t_1X_1+t_2X_2}(1) \\ &= \exp(\mu_1t_1 + \mu_2t_2 + \frac{1}{2}\sigma_1^2t_1^2 + \frac{1}{2}\sigma_2^2t_2^2 + \rho\sigma_1\sigma_2t_1t_2), \forall (t_1, t_2) \in \mathbb{R}^2. \end{aligned}$$

Since an MGF determines the distribution, we conclude $X = (X_1, X_2) \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

Remark 1.386 (Interpretation of parameters appearing in the p.d.f. of a Continuous RV). In the examples of continuous RVs discussed in this course, we have seen that certain parameters appear in the description of p.d.fs. If we specify the values of these parameters, then we obtain a specific example of distribution from a family of possible distributions. In certain cases, we have already been able to interpret them in terms of properties of the distribution of the RV. For example, if $X \sim N(\mu, \sigma^2)$, then $\mu = \mathbb{E}X$ and $\sigma^2 = Var(X)$. We list some interpretation of these parameters.

(a) (Location parameter) If we have a family of p.d.f.s $f_\theta, \theta \in \Theta$, where θ is a real valued parameter (i.e. $\Theta \subseteq \mathbb{R}$) and if $f_\theta(x) = f_0(x - \theta), \forall x \in \mathbb{R}$, then we say that θ is a location parameter for the family of distributions given by the p.d.f.s f_θ. In this case, the family is called a location family and the p.d.f. f_0 is free of θ, i.e. does not depend on θ. We can restate this fact in terms of the corresponding RVs X_θ as follows: the p.d.f./distribution of $X_\theta - \theta$ does not depend on θ.

(b) (Scale parameter) If we have a family of p.d.f.s f_θ, where θ is a real valued parameter (i.e. $\Theta \subseteq \mathbb{R}$) and if $f_\theta(x) = \frac{1}{\theta}f_1(\frac{x}{\theta}), \forall x \in \mathbb{R}$, then we say that θ is a scale parameter for the family of distributions given by the p.d.f.s f_θ. In this case, the family is called a scale family and the p.d.f. f_1 is free of θ, i.e. does not depend on θ. We can restate this fact in terms of the corresponding RVs X_θ as follows: the p.d.f./distribution of $\frac{1}{\theta}X_\theta$ does not depend on θ.

(c) (Location-scale parameter) If we have a family of p.d.f.s $f_{\mu,\sigma}$ with $\sigma > 0$ and if $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right) = f_{0,1}(x), \forall x \in \mathbb{R}$, then we say that (μ, σ) is a location-scale parameter for the family of distributions given by the p.d.f.s $f_{\mu,\sigma}$. In this case, the family is called a location-scale family and the p.d.f. $f_{0,1}$ is free of (μ, σ), i.e. does not depend on (μ, σ). We can restate this fact in terms of the corresponding RVs $X_{\mu,\sigma}$ as follows: the p.d.f./distribution of $\frac{X_{\mu,\sigma}-\mu}{\sigma}$ does not depend on (μ, σ).

(d) (Shape parameter) Some family of p.d.f.s also has a shape parameter, where changing the value of the parameter affects the shape of the graph of the p.d.f..

Example 1.387. (a) The family of RVs $X_{\mu,\theta} \sim \mathit{Cauchy}(\mu, \theta), \mu \in \mathbb{R}, \theta > 0$ with the p.d.f.

$$f_{\mu,\theta}(x) = \frac{\theta}{\pi}\frac{1}{\theta^2 + (x-\mu)^2}, \forall x \in \mathbb{R}$$

is a location-scale family with location parameter μ and scale parameter θ.

(b) For the family of RVs $X_\alpha \sim \mathit{Gamma}(\alpha, 1), \alpha > 0$ with the p.d.f.

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

α is a shape parameter.

Definition 1.388 (Weibull distribution)**.** We say that an RV X follows the Weibull distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$, if its p.d.f. is given by

$$f_X(x) = \begin{cases} \frac{\alpha}{\beta^\alpha}x^{\alpha-1}\exp\left[-\left(\frac{x}{\beta}\right)^\alpha\right], & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note 1.389. Let $X \sim \mathit{Exponential}(\beta^\alpha)$ for some $\alpha, \beta > 0$. Then $Y = X^{\frac{1}{\alpha}}$ follows the Weibull distribution with shape parameter α and scale parameter β.

Definition 1.390 (Pareto distribution)**.** We say that an RV X follows the Pareto distribution with scale parameter $\theta > 0$ and shape parameter $\alpha > 0$, if its p.d.f. is given by

$$f_X(x) = \begin{cases} \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Remark 1.391 (Descriptive Measures of Probability Distributions)*.* The distribution of an RV provides numerical values through which we can quantify/understand the manner in which the RV takes values in various subsets of the real line. However, at times, it is difficult to grasp the features of the RV from the distribution. As an alternative, we typically use four types of numerical quantities associated with the distribution to summarize the information. We refer to them as descriptive measures of the probability distribution.

(a) Measures of Central Tendency or location: here, we try to find a 'central' value around which the possible values of the RV are distributed.

(b) Measures of Dispersion: once we have an idea of the 'central' value of the RV (equivalently, the probability distribution), we check the scattering/dispersion of the all the possible values of the RV around this 'central' value.

(c) Measures of Skewness: here, we try to quantify the asymmetry of the probability distribution.

(d) Measures of Kurtosis: here, we try to measure the thickness of the tails of the RV (equivalently, the probability distribution) while comparing with the Normal distribution.

We describe these measures along with examples.

Example 1.392 (Measures of Central Tendency)**.** (a) The Mean of an RV is a good example of a measure of central tendency. It also has the useful property of linearity. However, it may be affected by few extreme values, referred to as the outliers. The mean may not exist for all distributions.

(b) Median, i.e. a quantile of order $\frac{1}{2}$ of an RV is always defined and is usually not affected by a few outliers. However, the median lacks the linearity property, i.e. a median of $X + Y$ has no general relationship with the medians of X and Y. Further, a median focuses on the probabilities with which the values of the RV occur rather than the exact numerical values. A median need not be unique.

(c) The mode m_0 of a probability distribution is the value that occurs with 'highest probability', and is defined by $f_X(m_0) = \sup\{f_X(x) : x \in S_X\}$, where f_X denotes the p.m.f./p.d.f. of X, as appropriate and S_X denotes the support of X. Mode need not be unique. Distributions with one, two or multiple modes are called unimodal, bimodal or multimodal distributions, respectively. Usually, it is easy calculate. However, it may so happen that a distribution has more than multiple modes situated far apart, in which case it may not be suitable for a measure of central tendency.

Example 1.393 (Measures of Dispersion). (a) If the support S_X of an RV X is contained in the interval $[a, b]$ and this is the smallest such interval, then we define $b - a$ to be the range of X. This measure of dispersion does not take into account the probabilities with which the values of X are distributed.

(b) Mean Deviation about a point $c \in \mathbb{R}$: If $\mathbb{E}|X - c|$ exists, we define it to be the mean deviation of X about the point c. Usually, we take c to be the mean (if it exists) or the median and obtain mean deviation about the mean or median, respectively. However, it may be difficult to compute and even may not exist. The mean deviations are also affected by a few outliers.

(c) Standard Deviation: As defined earlier, the standard deviation of an RV X is $\sqrt{Var(X)}$, if it exists. Compared to the mean deviation, the standard deviation is usually easier to compute. The standard deviation is affected by a few outliers.

(d) Quartile Deviation: Recall that $\mathfrak{z}_{0.25}$ and $\mathfrak{z}_{0.75}$ denotes the lower and upper quartiles. We define $\mathfrak{z}_{0.75} - \mathfrak{z}_{0.25}$ to be the inter-quartile range and refer to $\frac{1}{2}[\mathfrak{z}_{0.75} - \mathfrak{z}_{0.25}]$ as the semi-inter-quartile range or the quartile deviation. This measures the spread in the middle half of the distribution and is therefore not influenced by extreme values. However, it does not take into account the numerical values of the RV.

(e) Coefficient of Variation: The coefficient of variation of X is defined as $\frac{\sqrt{Var(X)}}{\mathbb{E}X}$, provided $\mathbb{E}X \neq 0$. This aims to measure the variation per unit of mean. It, by definition, does not depend on the unit of measurement. However, it may be sensitive to small changes in the mean, if it is close to zero.

Note 1.394 (A Measure of Skewness)**.** If the distribution of an RV X is symmetric about the mean μ, then $f_X(\mu+x) = f_X(\mu - x), \forall x \in \mathbb{R}$, where f_X denotes the p.m.f./p.d.f. of X. If this is not the case, then two cases may occur.

(a) (Positively skewed) the distribution may have more probability mass towards the right hand side of the graph of f_X. In this case, the tails on the right hand side are longer.

(b) (Negatively skewed) the distribution may have more probability mass towards the left hand side of the graph of f_X. In this case, the tails on the left hand side are longer.

To measure this asymmetry, we usually look at $\mathbb{E}Z^3$, where $Z = \frac{X-\mathbb{E}X}{\sqrt{Var(X)}}$, provided the moments exist. Note that Z is independent of the units of measurement and

$$\mathbb{E}Z^3 = \frac{\mathbb{E}(X-\mathbb{E}X)^3}{(Var(X))^{\frac{3}{2}}} = \frac{\mu_3(X)}{(\mu_2(X))^{3/2}}.$$

We may refer to a distribution being positively or negatively skewed according as the above quantity being positive or negative. If $X \sim Exponential(\lambda)$, then $\mathbb{E}Z^3 = 2$ and hence the distribution of X is positively skewed.

Note 1.395. There are many other measures of skewness used in practice. However, we do not discuss them in this course.

Note 1.396 (A measure of Kurtosis)**.** The probability distribution of X is said to have higher (respectively, lower) kurtosis than the Normal distribution, if its p.m.f./p.d.f., in comparison with the p.d.f. of a Normal distribution, has a sharper (respectively, rounded) peak and longer/fatter (respectively, shorter/thinner) tails. To measure the kurtosis of X, we look at $\mathbb{E}Z^4$, where $Z = \frac{X-\mathbb{E}X}{\sqrt{Var(X)}}$, provided the moments exist. Note that Z is independent of the units of measurement and

$$\mathbb{E}Z^4 = \frac{\mathbb{E}(X-\mathbb{E}X)^4}{(Var(X))^2} = \frac{\mu_4(X)}{(\mu_2(X))^2}.$$

If $X \sim N(\mu, \sigma^2)$, then $Z \sim N(0,1)$ and hence $\mathbb{E}Z^4 = 3$ (see Remark 1.248). For a general RV X, the quantity $\frac{\mu_4(X)}{(\mu_2(X))^2} - 3$ is referred to as the excess kurtosis of X. If the excess kurtosis is zero, positive or negative, then we refer to the corresponding probability distribution as mesokurtic, leptokurtic or platykurtic, respectively. If $X \sim Exponential(\lambda)$, then $\mathbb{E}Z^4 = 9$ and hence the distribution of X is leptokurtic.

Definition 1.397 (Quantile function of an RV)**.** Let X be an RV with the DF F_X. The function $Q_X : (0,1) \to \mathbb{R}$ defined by

$$Q_X(p) := \inf\{x \in \mathbb{R} : F_X(x) \geq p\}, \forall p \in (0,1)$$

is called the quantile function of X.

Proposition 1.398 (Probability integral transform)**.** *Let X be a continuous RV with the DF F_X, p.d.f. f_X and quantile function Q_X.*

(a) We have $F_X(X) \sim Uniform(0,1)$.

(b) For any $U \sim Uniform(0,1)$, we have $Q_X(U) \overset{d}{=} X$.

Proof. We prove only the first statement. The proof of the second statement is similar. Take $Y = F_X(X)$. Then,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(F_X(X) \leq y) = \begin{cases} 0, & \text{if } y < 0, \\ 1, & \text{if } y \geq 1. \end{cases}$$

For $y \in [0,1)$, we have

$$\mathbb{P}(F_X(X) = y) = \mathbb{P}(x_1 \leq X \leq x_2) = 0$$

for some $x_1, x_2 \in \mathbb{R}$ with $F_X(x_1) = F_X(x_2)$. Here, we have used the fact that X is a continuous RV. Now, for $y \in [0,1)$,

$$\begin{aligned} \mathbb{P}(F_X(X) \leq y) &= \mathbb{P}(F_X(X) < y) \\ &= 1 - \mathbb{P}(F_X(X) \geq y) \\ &= 1 - \mathbb{P}(X \geq Q_X(y)) \\ &= 1 - \mathbb{P}(X > Q_X(y)) \end{aligned}$$

$$\begin{aligned}&= \mathbb{P}(X \leq Q_X(y))\\ &= F_X(Q_X(y)\\ &= y.\end{aligned}$$

Hence, $Y = F_X(X) \sim Uniform(0, 1)$. This completes the proof. □

Note 1.399. Let X be an RV with the quantile function Q_X. If we can generate random samples $U_1, U_2, \cdots, U_n$ from $U \sim Uniform(0, 1)$, then $Q_X(U_1), Q_X(U_2), \cdots, Q_X(U_n)$ are random samples from the distribution of X. This observation may be used in practice to generate random samples for known distributions from the $Uniform(0, 1)$ distribution.

Note 1.400 (Moments do not determine the distribution of an RV)**.** Let $X \sim N(0, 1)$ and consider $Y = e^X$. The distribution of Y is usually called the lognormal distribution, since $\ln Y = X \sim N(0, 1)$. Using standard techniques, we can compute the p.d.f. of Y:

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} y^{-1} \exp\left[-\frac{(\ln y)^2}{2}\right], & \text{if } y > 0,\\ 0, & \text{otherwise.}\end{cases}$$

It can be shown that the continuous RVs $X_\alpha, \alpha \in [-1, 1]$ with the p.d.fs

$$f_{X_\alpha}(y) = f_Y(y) \left[1 + \alpha \sin(2\pi \ln y)\right], \forall y \in \mathbb{R}$$

has the same moments as Y. However, the distributions are different. This shows that the moments of an RV do not determine the distribution. (see the article ‘On a property of the lognormal distribution’ by C.C. Heyde, published in Journal of the Royal Statistical Society: Series B, volume 29 (1963).)

Note 1.401 (Operations on DFs)**.** Recall that a DF $F : \mathbb{R} \to [0, 1]$ is characterized by the properties that it is right continuous, non-decreasing and $\lim_{x\to\infty} F(x) = 1, \lim_{x\to-\infty} F(x) = 0$. Given two DFs $F, G : \mathbb{R} \to [0, 1]$ and $\alpha \in [0, 1]$, we make the following observations.

(a) (Convex combination of DFs) The function $H : \mathbb{R} \to [0, 1]$ defined by $H(x) := \alpha F(x) + (1 - \alpha)G(x), \forall x \in \mathbb{R}$ has the relevant properties and hence is a DF.

(b) (Product of DFs) The function $H : \mathbb{R} \to [0,1]$ defined by $H(x) := F(x)G(x), \forall x \in \mathbb{R}$ has the relevant properties and hence is a DF. In particular, F^2 is a DF, if F is so.

In fact, a general DF can be written as a convex combination of discrete DFs and some special continuous DFs. We do not discuss such results in this course.

Remark 1.402. In practice, given a known RV X, many times we need to find out the distribution of $h(X)$ for some function $h : \mathbb{R} \to \mathbb{R}$ or even, simply, compute the expectations of the form $\mathbb{E}h(X)$. As already discussed earlier in the course, we can theoretically (i.e., in principle) compute $\mathbb{E}h(X)$ as $\int_{-\infty}^{\infty} h(x) f_X(x)\, dx$, when X is a continuous RV with p.d.f. f_X, for example. However, in practice, it may happen that this integral does not have a closed form expression – which makes it challenging to evaluate. The problem becomes more intractable when we look at similar problems where X is a random vector and the joint/marginal distributions need to be considered. In such situations, as an alternative, we try to find 'good' approximations for the quantities of interest, where the approximation terms are easier to compute than the original expression. This motivation leads to the various notions for convergence of RVs. If some quantity of interest involving an RV X, say $\mathbb{E}X$, is difficult to compute, then we find an appropriate 'approximating' sequence of RVs $\{X_n\}_n$ for X and use the values $\mathbb{E}X_n$ as an approximation for $\mathbb{E}X$.

Remark 1.403. Given a random sample $X_1, X_2, \cdots, X_n$ from $N(\mu, \sigma^2)$ distribution, consider the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Here, we have written $\bar{X}_n$, instead of just $\bar{X}$, to highlight the dependence of the sample mean on the sample size n. Recall that $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. The behaviour of $\bar{X}_n$ for large n is of interest. This is also another motivation for us to study the convergence of sequences of RVs.

We now discuss concepts for convergence of sequences of RVs.

Definition 1.404 (Convergence in r-th mean)**.** Let $X, X_1, X_2, \cdots$ be RVs defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $r \geq 1$. If $\mathbb{E}|X|^r < \infty, \mathbb{E}|X_n|^r < \infty, \forall n$ and if

$$\lim_{n \to \infty} \mathbb{E}|X_n - X|^r = 0,$$

then we say that the sequence $\{X_n\}_n$ converges to X in r-th mean.

Note 1.405. (a) If a sequence $\{X_n\}_n$ converges to X in r-th mean for some $r \geq 1$, then we have

$$\lim_{n\to\infty} \mathbb{E}|X_n|^r = \mathbb{E}|X|^r,$$

and

$$\lim_{n\to\infty} \mathbb{E}X_n^r = \mathbb{E}X^r,$$

i.e., we have the convergence of the r-th moments.

(b) The sequence $\{X_n\}_n$ converges to X in r-th mean if and only if the sequence $\{X_n - X\}_n$ converges to 0 in r-th mean.

Remark 1.406. Even though we have defined the r-th order moments for $0 < r < 1$, for technical reasons we do not consider the convergence in r-th mean in this case. The details are beyond the scope of this course. In what follows, whenever we consider the convergence in r-th mean, we assume $r \geq 1$.

Definition 1.407 (Convergence in Probability)**.** Let $X, X_1, X_2, \cdots$ be RVs defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If for all $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0,$$

then we say that the sequence $\{X_n\}_n$ converges to X in probability and write $X_n \xrightarrow[n\to\infty]{P} X$.

Note 1.408. (a) Suppose that a sequence $\{X_n\}_n$ converges to X in probability. Now, for all $\epsilon > 0$, note that

$$\mathbb{P}(|X_n - X| \geq 2\epsilon) \leq \mathbb{P}(|X_n - X| > \epsilon) \leq \mathbb{P}(|X_n - X| \geq \epsilon).$$

Convergence in probability is equivalent to the fact that

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

(b) The sequence $\{X_n\}_n$ converges to X in probability if and only if the sequence $\{X_n - X\}_n$ converges to 0 in probability.

Proposition 1.409. *Let $X, X_1, X_2, \cdots$ be RVs defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If the sequence $\{X_n\}_n$ converges to X in r-th mean for some $r \geq 1$, then $X_n \xrightarrow[n\to\infty]{P} X$.*

Proof. By Markov's inequality (Corollary 1.255), we have

$$\mathbb{P}(|X_n - X| > \epsilon) \leq \epsilon^{-r}\, \mathbb{E}|X_n - X|^r.$$

Since $\lim_{n\to\infty} \mathbb{E}|X_n - X|^r = 0$, we have the result. □

Corollary 1.410. *Let $\{X_n\}_n$ be a sequence of RVs with finite second moments. If $\lim_n \mathbb{E}X_n = \mu$ and $\lim_n Var(X_n) = 0$, then $\{X_n\}_n$ converges to μ in 2nd mean and in particular, in probability.*

Proof. We have $\mathbb{E}|X_n - \mu|^2 = \mathbb{E}\left[(X_n - \mu_n) + (\mu_n - \mu)\right]^2 = \mathbb{E}(X_n - \mu_n)^2 + (\mu_n - \mu)^2 = Var(X_n) + (\mu_n - \mu)^2$. By our hypothesis, $\lim_n \mathbb{E}|X_n - \mu|^2 = 0$. Hence, $\{X_n\}_n$ converges to μ in 2nd mean. By Proposition 1.409, the sequence also converges in probability. □

Example 1.411. Let $X_1, X_2, \cdots$ be i.i.d. $Uniform(0, \theta)$ RVs, for some $\theta > 0$. The sequence $\{X_n\}_n$ being i.i.d. means that the collection $\{X_n : n \geq 1\}$ is mutually independent and that all the RVs have the same law/distribution. Here, the common p.d.f. and the common DF are given by

$$f(x) = \begin{cases} \frac{1}{\theta}, & \text{if } x \in (0, \theta), \\ 0, & \text{otherwise} \end{cases}, \quad F(x) = \begin{cases} 0, & \text{if } x < 0, \\ \frac{x}{\theta}, & \text{if } 0 \leq x < \theta, \\ 1, & \text{if } x \geq \theta. \end{cases}$$

Consider $X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$. Using Proposition 1.344, we have the marginal p.d.f. of $X_{(n)}$ is given by

$$g_{X_{(n)}}(x) = \begin{cases} \frac{n}{\theta^n} x^{n-1}, & \text{if } x \in (0, \theta), \\ 0, & \text{otherwise.} \end{cases}$$

Then,

$$\mathbb{E}X_{(n)} = \int_0^\theta x\, \frac{n}{\theta^n} x^{n-1}\, dx = \frac{n}{n+1}\theta, \quad \mathbb{E}X_{(n)}^2 = \int_0^\theta x^2\, \frac{n}{\theta^n} x^{n-1}\, dx = \frac{n}{n+2}\theta^2$$

and

$$Var(X_{(n)}) = \theta^2 \left[\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right] = \theta^2 \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} = \theta^2 \frac{n}{(n+2)(n+1)^2}.$$

Now, $\lim_n \mathbb{E}X_{(n)} = \theta$ and $\lim_n Var(X_{(n)}) = 0$. Hence, by Corollary 1.410, $\{X_{(n)}\}_n$ converges in 2nd mean to θ and also in probability.

Remark 1.412 (Convergence in probability does not imply convergence in r-th mean). Consider a sequence of discrete RVs $\{X_n\}_n$ with $X_n \sim Bernoulli(\frac{1}{n}), \forall n$. Consider $Y_n := nX_n, \forall n$. Then Y_n's are also discrete with the p.m.f.s given by

$$f_{Y_n}(y) = \begin{cases} 1 - \frac{1}{n}, \text{ if } y = 0, \\ \frac{1}{n}, \text{ if } y = n, \\ 0, \text{ otherwise.} \end{cases}$$

For all $\epsilon > 0$, we have $\mathbb{P}(|Y_n| \geq \epsilon) = \frac{1}{n} \xrightarrow{n\to\infty} 0$ and hence $Y_n \xrightarrow[n\to\infty]{P} 0$. But, for any $r > 1$, $\mathbb{E}|Y_n|^r = n^{r-1}, \forall n$. Here, $\{Y_n\}_n$ does not converge to 0 in r-th mean.

Example 1.413. $X_1, X_2, \cdots$ be i.i.d. RVs following $N(\mu, \sigma^2)$ distribution. Recall that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Then $\lim_n \mathbb{E}\bar{X}_n = \lim_n \mu = \mu$ and $\lim_n Var(\bar{X}_n) = \lim_n \frac{\sigma^2}{n} = 0$. By Corollary 1.410, $\{\bar{X}_n\}_n$ converges in 2nd mean to μ and also in probability.

The above example leads to the following result.

Theorem 1.414 (Weak Law of Large Numbers (WLLN))**.** *Let $X_1, X_2, \cdots$ be i.i.d. RVs such that $\mathbb{E}X_1$ exists. Then, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{P} \mathbb{E}X_1$.*

Remark 1.415. We only discuss the proof of Theorem 1.414, when $\mathbb{E}X_1^2$ exists. The proof of the theorem when $\mathbb{E}X_1^2$ does not exist is beyond the scope of this course. However, we shall use this theorem in its full generality.

Proof of WLLN (Theorem 1.414) (assuming $\mathbb{E}X_1^2 < \infty$). Observe that $\mathbb{E}\bar{X}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i = \frac{1}{n}n\mathbb{E}X_1 = \mathbb{E}X_1$ and, using independence of X_i's we have

$$Var(\bar{X}_n) = \frac{1}{n^2}Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}Var(X_1) \xrightarrow{n\to\infty} 0.$$

By Corollary 1.410, the result follows. □

Remark 1.416. The WLLN suggests that for large sample sizes, the sample mean based on a random sample from a given distribution (also referred to as a population) is close to the expectation/mean of the distribution, in the sense of convergence in probability. In practice, this principle can be used to find an approximate value of the expectation of a distribution.

Example 1.417. Let $\{X_n\}_n$ be i.i.d. RVs with the common distribution $Bernoulli(p)$ for some $p > 0$. Here, we may visualize X_n's as a sequence of coin tosses with probability of success (obtaining head) as p. By the WLLN, $\bar{X}_n \xrightarrow[n\to\infty]{P} \mathbb{E}X_1 = p$, i.e. for all $\epsilon > 0$, $\lim_{n\to\infty}\mathbb{P}(|\bar{X}_n - p| \geq \epsilon) = 0$. This supports the intuitive notion that by tossing a coin, with unknown p, a large number of times we can make an educated guess about the value of p.

Example 1.418. Continuing with the discussion of the previous example, we can justify the working methodology of assigning probabilities by a relative frequency approach. Suppose we repeat a random experiment n times and observe whether an event E occurs or not in each trial. For $i = 1, 2, \cdots, n$, we consider an RV X_i to be 1 if E occurs and 0 otherwise. As discussed earlier in Remark 1.228, $X_i \sim Bernoulli(p)$, where $p = \mathbb{P}(E)$. If p is unknown, then by the WLLN we have $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{P} p$, i.e., the observed relative frequency $\frac{1}{n}\sum_{i=1}^n X_i$ in first n trials approximates p in probability, for large n.

Note 1.419. There is a stronger version of the WLLN, called the Strong Law of Large Numbers, which we do not discuss in the course.

Theorem 1.420 (Continuous Mapping Theorem for convergence in Probability)**.** *Let $\{X_n\}_n$ be a sequence of RVs converging to X in probability. Let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then $\{h(X_n)\}_n$ converges to $h(X)$ in probability.*

Note 1.421. In general, continuous mapping theorem is not true for convergence in r-th mean. Construction of such an example is left as an exercise in the problem set 10.

The proof of the next result is not part of the course.

Theorem 1.422 (Algebraic Operations with Convergence in Probability)**.** *Let $\{X_n\}_n$ and $\{Y_n\}$ be sequences of RVs such that $X_n \xrightarrow[n\to\infty]{P} x$ and $Y_n \xrightarrow[n\to\infty]{P} y$ for some $x, y \in \mathbb{R}$. Let $\{a_n\}_n$ and $\{b_n\}_n$ be sequences in $\mathbb{R}$ converging to $a, b \in \mathbb{R}$ respectively. Then the following statements hold.*

(a) $X_n + Y_n \xrightarrow[n\to\infty]{P} x + y$.

(b) $X_n - Y_n \xrightarrow[n\to\infty]{P} x - y$.

(c) $X_nY_n \xrightarrow[n\to\infty]{P} xy$.

(d) $\frac{X_n}{Y_n} \xrightarrow[n\to\infty]{P} \frac{x}{y}$, *provided* $y \neq 0$.

(e) $a_nX_n + b_n \xrightarrow[n\to\infty]{P} ax + b$.

Remark 1.423. Let $X_1, X_2, \cdots$ be i.i.d. RVs such that $\mathbb{E}X_1^2$ exists. Consider the sample variance $S_n^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}(\bar{X}_n)^2$. By the assumption, the RVs X_i^2 are i.i.d. with finite expectation $\mathbb{E}X_1^2$, and hence by the WLLN (Theorem 1.414)

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow[n\to\infty]{P} \mathbb{E}X_1^2.$$

Again by WLLN $\bar{X}_n \xrightarrow[n\to\infty]{P} \mathbb{E}X_1$ and by Theorem 1.420 applied to the function $h(x) = x^2, \forall x \in \mathbb{R}$, we have

$$(\bar{X}_n)^2 \xrightarrow[n\to\infty]{P} (\mathbb{E}X_1)^2.$$

Since $\frac{n}{n-1} \xrightarrow{n\to\infty} 1$, using Theorem 1.422, we have $S_n^2 \xrightarrow[n\to\infty]{P} Var(X_1)$. By Theorem 1.420, we have $S_n \xrightarrow[n\to\infty]{P} \sqrt{Var(X_1)}$.

Note 1.424. In the discussion involving convergence of RVs, we have seen two notions of convergence, viz. convergence in r-th mean and convergence in probability. Now, given a sequence of RVs $\{X_n\}_n$, the law/distribution of each X_n is determined by the corresponding DFs F_{X_n}. It is, therefore, reasonable to consider the problem of the convergence of the DFs.

Remark 1.425 (Pointwise limit of DFs need not be a DF). We show by examples that the pointwise limit of DFs need not be a DF.

(a) Let $X_n \sim Uniform(-n, n)\,\forall n = 1, 2, \cdots$. Here,

$$F_{X_n}(x) = \begin{cases} 0, \text{ if } x > -n, \\ \frac{x+n}{2n}, \text{ if } -n \leq x < n, \\ 1, \text{ if } x \geq n. \end{cases}$$

Then, the pointwise limit exists and is given by $\lim_n F_{X_n}(x) = \frac{1}{2}, \forall x$. However, the pointwise limit function, say, $F(x) = \frac{1}{2}, \forall x$ is not a DF.

(b) Consider the sequence $\{X_n\}_n$ with X_n degenerate at $\frac{1}{n}$. Then,

$$F_{X_n}(x) = \begin{cases} 0, \text{ if } x < \frac{1}{n}, \\ 1, \text{ if } x \geq \frac{1}{n}. \end{cases}$$

Then, the pointwise limit function exists and is given by

$$F(x) := \lim_n F_{X_n}(x) = \begin{cases} 0, \text{ if } x \leq 0, \\ 1, \text{ if } x > 0. \end{cases}$$

Since F is not right continuous at 0, it is not a DF. However, we may change the value of F at 0 and obtain the following DF $\widetilde{F}$ given by

$$\widetilde{F}(x) := \begin{cases} 0, \text{ if } x < 0, \\ 1, \text{ if } x \geq 0. \end{cases}$$

and $\widetilde{F}$ matches with $\lim_n F_{X_n}$ except at the point of discontinuity of $\widetilde{F}$. Note that $\widetilde{F}$ is the DF of the degenerate RV at 0.

Motivated by the above examples, we now consider the following notion of convergence of RVs.

Definition 1.426 (Convergence in Law/Distribution)**.** Let X be an RV defined on a Probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with DF F. For each n, let X_n be an RV defined on (possibly different) probability space $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ with DF F_n. Let D_F denote the point of discontinuities of F. We say the sequence $\{X_n\}_n$ converges in law/distribution to X, denoted by $X_n \xrightarrow[n\to\infty]{d} X$, if

$$F_n(x) \xrightarrow{n\to\infty} F(x), \forall x \in D_F^c.$$

Example 1.427. Consider the sequence $\{X_n\}_n$ with X_n degenerate at $\frac{1}{n}$ and let X be an RV degenerate at 0. Then, as discussed in Remark 1.425, we have $X_n \xrightarrow[n\to\infty]{d} X$.

Note 1.428. (a) Recall that the set D_F of discontinuities of a DF F, if it is non-empty, is either finite or countably infinite. If $X_n \xrightarrow[n\to\infty]{d} X$, then as per the definition, we must have $F_n(x) \xrightarrow{n\to\infty} F(x)$ everywhere, except possibly at a countable number of points.

(b) If X is a continuous RV, then $D_F = \emptyset$. If $X_n \xrightarrow[n\to\infty]{d} X$, then as per the definition $F_n(x) \xrightarrow{n\to\infty} F(x), \forall x \in \mathbb{R}$.

(c) Even if the RVs $X, X_1, X_2, \cdots$ are defined on different probability spaces, we can consider the notion of convergence in law/distribution. However, to consider the notion of convergence in r-th mean or in probability, we must have the RVs defined on the same probability space.

We state the following result without proof. The details of the proof are not part of the course.

Proposition 1.429. *Let the RVs $X, X_1, X_2, \cdots$ be defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $X_n \xrightarrow[n\to\infty]{P} X$ (or converges in the r-th mean for some $r \geq 1$), then $X_n \xrightarrow[n\to\infty]{d} X$.*

A special case of the above result requires more attention.

Proposition 1.430. *Let $\{X_n\}_n$ be a sequence of RVs defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $c \in \mathbb{R}$. Then $X_n \xrightarrow[n\to\infty]{P} c$ if and only if $X_n \xrightarrow[n\to\infty]{d} c$.*

Proof. Here, the constant c is being treated as a degenerate RV at c. The DF for this RV is given by

$$F(x) := \begin{cases} 0, & \text{if } x < c, \\ 1, & \text{if } x \geq c \end{cases}$$

with the only point of discontinuity at c. Now, $X_n \xrightarrow[n\to\infty]{d} c$ implies

$$\lim_n F_n(x) = F(x), \forall x \neq c,$$

where F_n denotes the DF of X_n. Observe that, for any $\epsilon > 0$,

$$\begin{aligned}\lim_n \mathbb{P}(|X_n - c| > \epsilon) &= \lim_n \mathbb{P}(X_n > c + \epsilon) + \lim_n \mathbb{P}(X_n < c - \epsilon)\\ &\leq \lim_n [1 - \mathbb{P}(X_n \leq c + \epsilon)] + \lim_n \mathbb{P}(X_n \leq c - \epsilon)\\ &= \lim_n [1 - F_n(c + \epsilon)] + \lim_n F_n(c - \epsilon)\\ &= [1 - F(c + \epsilon)] + F(c - \epsilon)\\ &= 0.\end{aligned}$$

This proves the sufficiency part. The necessity part follows from Proposition 1.429. □

Example 1.431. If a sequence of RVs converges in law/distribution, they need not converge in probability. Construction of such an example is left as an exercise in the problem set 10.

We now state some sufficient conditions which imply the convergence in law/distribution. The proofs are not included in the course.

Theorem 1.432. *Let $X, X_1, X_2, \cdots$ be RVs defined on the same probability space.*

(a) If these RVs are taking values in the set of non-negative integers (in particular, the RVs are discrete) and if the corresponding p.m.fs converge pointwise, i.e. $\lim_n f_{X_n}(x) = f(x), \forall x \in \{0, 1, 2, \cdots\}$, then $X_n \xrightarrow[n\to\infty]{d} X$.

(b) If all the RVs are continuous with the corresponding p.d.fs $f_X, f_{X_1}, f_{X_2}, \cdots$ and if $\lim_n f_{X_n}(x) = f(x), \forall x \in \mathbb{R}$, then $X_n \xrightarrow[n\to\infty]{d} X$.

(c) If these RVs have the MGFs $M, M_1, M_2, \cdots$ existing on $(-h, h)$ for some $h > 0$ and if $\lim_n M_n(t) = M(t), \forall t \in (-h, h)$, then $X_n \xrightarrow[n\to\infty]{d} X$.

Example 1.433. Consider the discrete RVs X_n with the p.m.f.s and DFs given by

$$f_{X_n}(x) = \begin{cases} \frac{1}{2}, \text{ if } x \in \{\frac{1}{2n}, \frac{1}{n}\}, \\ 0, \text{ otherwise} \end{cases}, \quad F_{X_n}(x) = \begin{cases} 0, \text{ if } x < \frac{1}{2n}, \\ \frac{1}{2}, \text{ if } \frac{1}{2n} \leq x < \frac{1}{n}, \\ 1, \text{ if } x \geq \frac{1}{n}. \end{cases}$$

Since

$$\lim_n F_{X_n}(x) = \begin{cases} 0, \text{ if } x < 0, \\ 1, \text{ if } x > 0 \end{cases}$$

equals the DF F of the degenerate RV at 0, except at the point of discontinuity 0 of F, we have $X_n \xrightarrow[n\to\infty]{d} 0$. However, $\lim_n f_{X_n}(0) = 0 \neq 1 = f_X(0)$. Here, the pointwise convergence of the p.m.f.s do not hold.

Example 1.434. Let $X, X_1, X_2, \cdots$ be independent RVs with $X \sim N(0,1)$ and $X_n \sim N(0, 1+\frac{1}{n})$. Looking at the MGFs we have

$$\lim_n M_{X_n}(t) = \lim_n \exp\left(\frac{1}{2}\left(1 + \frac{1}{n}\right)t^2\right) = \exp\left(\frac{1}{2}t^2\right) = M_X(t), \forall t \in \mathbb{R}.$$

Therefore, $X_n \xrightarrow[n\to\infty]{d} X$. However, using the independence of $X, X_1, X_2, \cdots$, we have $X_n - X \sim N(0, 2+\frac{1}{n})$ and an argument similar to above shows that $X_n - X \xrightarrow[n\to\infty]{d} Z$, where $Z \sim N(0,2)$. Here, $X_n - X$ does not converge to the degenerate RV at 0.

The proof of the next result is not included in the course.

Theorem 1.435 (Continuous Mapping Theorem for Convergence in Distribution)**.** *Let $X_n \xrightarrow[n\to\infty]{d} X$ and $Y_n \xrightarrow[n\to\infty]{P} c$ for some $c \in \mathbb{R}$.*

(a) Let $h : \mathbb{R} \to \mathbb{R}$ be a function continuous on the support S_X of X. Then $h(X_n) \xrightarrow[n\to\infty]{d} h(X)$.

(b) Let $h : \mathbb{R}^2 \to \mathbb{R}$ be a function continuous on the set $\{(x,y) \in \mathbb{R}^2 : x \in S_X, y = c\}$. Then $h(X_n, Y_n) \xrightarrow[n\to\infty]{d} h(X, c)$.

Notation 1.436. For any RV X, we treat $0 \times X$ as an RV degenerate at 0.

A special case of the above theorem is useful in practice.

Theorem 1.437 (Slutky's Theorem)**.** *Let* $X_n \xrightarrow[n\to\infty]{d} X$ *and* $Y_n \xrightarrow[n\to\infty]{P} c$ *for some* $c \in \mathbb{R}$*. Then* $X_n + Y_n \xrightarrow[n\to\infty]{d} X + c$ *and* $X_nY_n \xrightarrow[n\to\infty]{d} cX$*.*

Note 1.438. We now look at an example of convergence in distribution which is quite useful in practice.

Theorem 1.439 (Poisson approximation to Binomial Distribution)**.** *Let* $X_n \sim Binomial(n, p_n), n = 1, 2, \cdots$ *where* $p_n \in (0, 1), \forall n$ *and*

$$\lim_n np_n = \lambda > 0.$$

Then $X_n \xrightarrow[n\to\infty]{d} X$ *with* $X \sim Poisson(\lambda)$ *with* $\mathbb{P}(X_n = k) \xrightarrow{n\to\infty} \mathbb{P}(X = k)$ *for all* $k = 0, 1, 2, \cdots$*.*

Proof. We prove the stated convergence using MGFs (see Theorem 1.432). We have for all $t \in \mathbb{R}$

$$\lim_n M_{X_n}(t) = \lim_n (1 - p_n + p_n e^t)^n = \lim_n \left(1 + \frac{np_n(e^t - 1)}{n}\right)^n = \exp(\lambda(e^t - 1)) = M_X(t).$$

Hence, we have the convergence in law/distribution.
We may check the same using the p.m.fs. For fixed $k = 0, 1, 2, \cdots$ and for all $n \geq k$, we have

$$\begin{aligned}\lim_n \mathbb{P}(X_n = k) &= \lim_n \binom{n}{k} p_n^k (1 - p_n)^{n-k} \\ &= \frac{1}{k!} \lim_n \frac{n(n-1)\cdots(n-k+1)}{n^k} (np_n)^k \left(1 - \frac{np_n}{n}\right)^{n-k} \\ &= \frac{1}{k!} \lambda^k e^{-\lambda} \\ &= \mathbb{P}(X = k).\end{aligned}$$

This completes the proof. □

Remark 1.440. In Theorem 1.439, from the assumption $\lim_n np_n = \lambda > 0$, we have the probability of success p_n is 'small' for large n. We may therefore treat X_n as the number of successes of a 'rare' event in n trials of a random experiment with probability of success p_n (see Remark 1.351). Here, we have kept $\mathbb{E}X_n = np_n$ close to $\lambda > 0$. So the number n of trials are increases, but the probability of success is decreases with n.

Example 1.441. If $X \sim Binomial(1000, 0.003)$, then the exact value of

$$\mathbb{P}(X = 5) = \binom{1000}{5}(0.003)^5(0.997)^{995}$$

is hard to compute. Instead, we can approximate the value by $\mathbb{P}(Y = 5)$ where $Y \sim Poisson(1000\times 0.003) = Poisson(3)$. Here, $\mathbb{P}(Y = 5) = e^{-3}\,\frac{3^5}{5!}$ is comparatively easier to compute.

Remark 1.442 (A question about the rate of convergence). We have seen three types of convergences of RVs and their examples. However, in these examples, we can ask how 'fast' does the convergences occur? For example, by the WLLN (Theorem 1.414), we have $\bar{X}_n \xrightarrow[n\to\infty]{P} \mathbb{E}X_1$ for any i.i.d. sequence of RVs $X_1, X_2, \cdots$ with finite expectation. How 'fast' does the 'error term' $\bar{X}_n - \mathbb{E}X_1$ go to 0? In other words, how 'small' is $\bar{X}_n - \mathbb{E}X_1$ for 'large' n? If we can show, '$n^\alpha\left(\bar{X}_n - \mathbb{E}X_1\right) \xrightarrow{n\to\infty} c$' for some $c \in \mathbb{R}, \alpha > 0$, then for large n, we may say $\bar{X}_n - \mathbb{E}X_1$ is close to $\frac{c}{n^\alpha}$ - which gives an idea about the magnitude. This, however, is only an idea and not a concrete result. In fact, in this description, it is more likely to have an RV in the place of c above, with a clear notion of convergence for the 'error term', again in terms of some notion of convergence of RVs. It is to be noted that a convergence result with a 'rate of convergence' is stronger than another convergence result without any clear 'rate of convergence'. We shall come back to this discussion later.

Note 1.443. For i.i.d. RVs, recall from the proof of WLLN (Theorem 1.414), that $Var(\bar{X}_n) = \frac{1}{n^2}Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}Var(X_1)$. Provided $Var(X_1) > 0$, we have $\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{\sqrt{Var(X_1)}}$ is an RV with mean 0 and variance 1.

Theorem 1.444 (Central Limit Theorem (CLT))**.** *Let $X_1, X_2, \cdots$ be i.i.d. RVs such that $\mathbb{E}X_1^2$ exists and $Var(X_1) = \sigma^2 > 0$. Then,*

$$\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{\sigma} \xrightarrow[n\to\infty]{d} Z,$$

where $Z \sim N(0, 1)$.

Remark 1.445 (Restatements of the CLT). Under the assumptions of the CLT above, we can restate the conclusion in various useful ways. Note that the DF Φ of $Z \sim N(0, 1)$ is continuous everywhere on $\mathbb{R}$.

(a) $\lim_{n\to\infty} \mathbb{P}(\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{\sigma} \leq x) = \Phi(x), \forall x \in \mathbb{R}$.

(b) For all $a < b$, we have

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(a < \sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{\sigma} \leq b) \\
&= \lim_{n\to\infty} \mathbb{P}(\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{\sigma} \leq b) - \lim_{n\to\infty} \mathbb{P}(\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{\sigma} \leq a) \\
&= \Phi(b) - \Phi(a) \\
&= \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt
\end{aligned}$$

(c) Writing $Y_n = X_1 + X_2 + \cdots + X_n$, for all $a < b$ we have

$$\lim_{n\to\infty} \mathbb{P}(a < \frac{Y_n - n\mathbb{E}X_1}{\sigma\sqrt{n}} \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

Proof of CLT (Theorem 1.444). We find the limit of the MGFs of

$$Z_n := \sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mathbb{E}X_1}{\sigma}.$$

Here, $\frac{X_i - \mathbb{E}X_1}{\sigma}, i = 1, 2, \cdots$ are i.i.d. with mean 0 and variance 1. In particular, $\mathbb{E}\left(\frac{X_i - \mathbb{E}X_1}{\sigma}\right)^2 = 1$. Since we have assumed the existence of the MGFs, we have $M'_{\frac{X_i - \mathbb{E}X_1}{\sigma}}(0) = 0$ and $M''_{\frac{X_i - \mathbb{E}X_1}{\sigma}}(0) = 1$. We also have a Taylor series expansion in a neighbourhood of 0 as

$$M_{\frac{X_i - \mathbb{E}X_1}{\sigma}}(t) = M_{\frac{X_i - \mathbb{E}X_1}{\sigma}}(0) + tM'_{\frac{X_i - \mathbb{E}X_1}{\sigma}}(0) + \frac{t^2}{2}\left(M''_{\frac{X_i - \mathbb{E}X_1}{\sigma}}(0) + R(t)\right) = 1 + \frac{t^2}{2}(1 + R(t))$$

with $\lim_{t\to 0} R(t) = 0$.

Then, using the i.i.d. nature of $\frac{X_i - \mathbb{E}X_1}{\sigma}, i = 1, 2, \cdots$, we have

$$\begin{aligned}
M_{Z_n}(t) &= \mathbb{E}\exp\left(\frac{t}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mathbb{E}X_1}{\sigma}\right) \\
&= \left(M_{\frac{X_1 - \mathbb{E}X_1}{\sigma}}\left(\frac{t}{\sqrt{n}}\right)\right)^n \\
&= \left(1 + \frac{t^2}{2n}\left(1 + R\left(\frac{t}{\sqrt{n}}\right)\right)\right)^n
\end{aligned}$$

$$\xrightarrow{n\to\infty} \exp\left(\frac{t^2}{2}\right) = M_Z(t)$$

for all t in a neighbourhood of 0. Using Theorem 1.432, we conclude the proof. □

Remark 1.446. The CLT suggests that for large sample sizes, the normarlized version $\frac{\bar{X}_n - \mathbb{E}X_1}{\sqrt{Var(X_1)}}$ of the sample mean $\bar{X}_n$ based on a random sample from any given distribution is close to the Standard Normal distribution, in the sense of convergence in law/distribution. In practice, this result can be used to obtain estimates for probabilities involving the sample mean.

Note 1.447. In the statement of the CLT, we have taken $\sigma > 0$. If $\sigma = 0$, then note that all the RVs $X_1, X_2, \cdots$ are actually degenerate at some constant $c \in \mathbb{R}$ and $\bar{X}_n = c, \forall n$.

Remark 1.448 (From CLT to WLLN). Our motivation to study CLT type results was to find a 'rate of convergence' for the WLLN. As mentioned in Remark 1.442, a convergence result with a 'rate of convergence' is stronger than another convergence result without any clear 'rate of convergence'. We illustrate this idea by deriving the WLLN from the CLT. Under the assumptions of CLT (Theorem 1.444), we have

$$\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{\sigma} \xrightarrow[n\to\infty]{d} Z,$$

where $Z \sim N(0,1)$. Since $\frac{\sigma}{\sqrt{n}} \xrightarrow{n\to\infty} 0$, by Slutky's theorem (Theorem 1.437),

$$\bar{X}_n - \mathbb{E}X_1 \xrightarrow[n\to\infty]{d} 0 \times Z = Y,$$

where Y denotes an RV degenerate at 0. By Proposition 1.430, we have $\bar{X}_n - \mathbb{E}X_1 \xrightarrow[n\to\infty]{P} 0$. Finally, by Theorem 1.422, we conclude $\bar{X}_n \xrightarrow[n\to\infty]{P} \mathbb{E}X_1$, which is the WLLN. Note that, however, to show this we needed the additional assumption on the second moments, which is not required if we are only interested in the WLLN.

Note 1.449. As discussed above, using information from higher moments, we have improved results. The CLT stated here can be improved to the Berry-Esseen Theorem using information from 3-rd absolute moments and the WLLN can be improved to Hoeffding's inequality for bounded RVs. The CLT has a huge literature and many CLT-type results have been proved even in the non-i.i.d. setting. These results are not part of this course.

Remark 1.450. Let $X_1, X_2, \cdots$ be i.i.d. RVs such that $\mathbb{E}X_1^2$ exists and $Var(X_1) = \sigma^2 > 0$. By the CLT,

$$\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{\sigma} \xrightarrow[n\to\infty]{d} Z,$$

where $Z \sim N(0,1)$. From Remark 1.423 and Theorem 1.422, we have

$$\frac{\sigma}{S_n} \xrightarrow[n\to\infty]{P} 1,$$

where $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the sample variance. By Slutky's theorem,

$$\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X_1}{S_n} \xrightarrow[n\to\infty]{d} Z.$$

Note 1.451. Recall from Theorem 1.420 that if $X_n \xrightarrow[n\to\infty]{P} X$, then for any continuous function $h : \mathbb{R} \to \mathbb{R}$, we have $h(X_n) \xrightarrow[n\to\infty]{P} h(X)$. We can now ask about the rate of convergence. This question leads to a useful result, known as the Delta method. We do not discuss the proof of this result in this course.

Theorem 1.452 (Delta method). *Let $\{X_n\}_n$ be a sequence of RVs such that $n^b(X_n - a) \xrightarrow[n\to\infty]{d} X$ for $a \in \mathbb{R}, b > 0$ and some RV X. Let $g : \mathbb{R} \to \mathbb{R}$ be a function differentiable at a. Then*

$$n^b(g(X_n) - g(a)) \xrightarrow[n\to\infty]{d} g'(a)X.$$

Combining with the CLT, using the Delta method we get the following result often used in practice.

Theorem 1.453. *Let $X_1, X_2, \cdots$ be i.i.d. RVs such that $\mathbb{E}X_1^2$ exists and $Var(X_1) = \sigma^2 > 0$. Let $g : \mathbb{R} \to \mathbb{R}$ be a function differentiable at $a = \mathbb{E}X_1$ with $g'(a) \neq 0$. Then,*

$$\sqrt{n}\frac{g(\bar{X}_n) - g(\mathbb{E}X_1)}{\sigma} \xrightarrow[n\to\infty]{d} g'(a)Z \sim N(0, (g'(a))^2),$$

where $Z \sim N(0,1)$.

Remark 1.454. Let $X_1, X_2, \cdots$ be i.i.d. $Uniform(0,\theta)$ RVs, for some $\theta > 0$. Recall from Example 1.411 that $X_{(n)} = \max\{X_1, X_2, \cdots, X_n\} \xrightarrow[n\to\infty]{P} \theta$. We can now ask about the limiting distribution of $(\theta - X_{(n)})$ to understand the rate of convergence. Recall that the p.d.f. of $X_{(n)}$ is given by

$$g_{X_{(n)}}(x) = \begin{cases} \frac{n}{\theta^n} x^{n-1}, \text{ if } x \in (0,\theta), \\ 0, \text{ otherwise.} \end{cases}$$

Look at $Y_n := n(\theta - X_{(n)})$. Then for all $y \in \mathbb{R}$,

$$\begin{aligned} F_{Y_n}(y) &= \mathbb{P}(Y_n \leq y) \\ &= \mathbb{P}\left(X_{(n)} \geq \theta - \frac{y}{n}\right) \\ &= \int_{\theta - \frac{y}{n}}^{\infty} g_{X_{(n)}}(x)\, dx \\ &= \begin{cases} 0, \text{ if } y \leq 0 \\ 1 - \left(1 - \frac{y}{n\theta}\right)^n, \text{ if } 0 < y < n\theta, \\ 1, \text{ if } y > n\theta \end{cases} \\ &\xrightarrow{n\to\infty} \begin{cases} 0, \text{ if } y \leq 0 \\ 1 - \exp\left(-\frac{y}{\theta}\right), \text{ if } y > 0 \end{cases} \\ &= F_Y(y) \end{aligned}$$

where $Y \sim \mathit{Exponential}(\theta)$. Since the DF F_Y of Y is continuous everywhere, from the above computation we conclude that $Y_n = n(\theta - X_{(n)}) \xrightarrow[n\to\infty]{d} Y \sim \mathit{Exponential}(\theta)$. The sequence $\{X_{(n)}\}_n$ another example where Delta method can be applied.

2. To add

14.31 Let $X_1, X_2, \cdots, X_n$ be a random sample from $N(\mu, \sigma^2)$ distribution with the parameters μ and σ^2 being unknown. Recall from Remark 1.375 that $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$, where $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Then, $\mathbb{E}_{\mu,\sigma^2}\frac{(n-1)S_n^2}{\sigma^2} = (n-1)$ or $\mathbb{E}_{\mu,\sigma^2} S_n^2 = \sigma^2$, for all possible values of μ and σ^2.

14.39 Let $X_1, X_2, \cdots, X_n$ be a random sample from $N(\mu, \sigma^2)$ distribution, with the parameters μ and σ^2 being unknown. Consider the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Recall from Remark 1.369

that $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. By an application of the Chebyshev's inequality (Corollary 1.257), for all $\epsilon > 0$, we have

$$\mathbb{P}_{\mu,\sigma^2}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{1}{\epsilon^2} Var(\bar{X}_n) = \frac{1}{n\epsilon^2}\sigma^2$$

and hence $\lim_{n\to\infty} \mathbb{P}_{\mu,\sigma^2}(|\bar{X}_n - \mu| \geq \epsilon) = 0$ (also see Proposition 1.409). Therefore, $\{\bar{X}_n\}_n$ is consistent for μ.

15.10 We shall use the following pivotal quantities in our discussion. Let $X_1, X_2, \cdots, X_n$ be a random sample from $N(\mu, \sigma^2)$ distribution with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

(a) Suppose that μ is unknown but σ^2 is known. If the estimand is $h(\mu) = \mu$, then we consider the pivotal quantity $g(X_1, X_2, \cdots, X_n; \mu) = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim N(0,1)$ (see Remark 1.369).

(b) Suppose that μ and σ^2 both are unknown. If the estimand is $h(\mu, \sigma^2) = \mu$, then we consider the pivotal quantity $g(X_1, X_2, \cdots, X_n; \mu) = \sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1}$ (see Note 1.378).

(c) Suppose that σ^2 is unknown but μ is known. If the estimand is $h(\sigma^2) = \sigma^2$, then we consider the pivotal quantity $g(X_1, X_2, \cdots, X_n; \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2_n$ (see Note 1.372).

(d) Suppose that μ and σ^2 both are unknown. If the estimand is $h(\sigma^2) = \sigma^2$, then we consider the pivotal quantity $g(X_1, X_2, \cdots, X_n; \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \chi^2_{n-1}$ (see Remark 1.375).

16.1 (pivotal quantity – pooled sample mean)

17.11 If $Z \sim N(0,1)$, it can be checked that $\mathbb{P}(|Z| \leq 3) \approx 0.997$ and $\mathbb{P}(|Z| \leq 6) \approx 0.9997$. More generally, for $X \sim N(\mu, \sigma^2)$, we have $\mathbb{P}(|X - \mu| \leq 3\sigma) \approx 0.997$ and $\mathbb{P}(|X - \mu| \leq 6\sigma) \approx 0.9997$. This shows that the values of a normal RV is quite concentrated near its mean.

3. To add from practice problems

Set 9 question 4 (sample mean and sample variance)

Let $X_1, X_2, \cdots, X_n$ be a random sample from $N(\mu, \sigma^2)$ distribution. Consider the sample mean $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ and sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Show that $\bar{X}$ and S_n^2 are independent.

Look at the joint MGF of $(X_1 - \bar{X}, X_2 - \bar{X}, \cdots, X_n - \bar{X}, \bar{X})$ given by

$$M(t_1, t_2, \cdots, t_n, t_{n+1}) = \mathbb{E}\exp\left(\sum_{j=1}^n t_j(X_j - \bar{X}) + t_{n+1}\bar{X}\right), \forall (t_1, t_2, \cdots, t_n, t_{n+1}) \in \mathbb{R}^{n+1}$$

$$= \mathbb{E}\exp\left(\sum_{j=1}^{n} s_j X_j\right),$$

where $s_j = t_j + \frac{t_{n+1}-\sum_{i=1}^{n} t_i}{n}$. Using the independence of X_j's, we have

$$\begin{aligned}
M(t_1, t_2, \cdots, t_n, t_{n+1}) &= \prod_{j=1}^{n} \mathbb{E}\exp(s_j X_j) \\
&= \prod_{j=1}^{n} \exp\left(\mu s_j + \frac{1}{2}\sigma^2 s_j^2\right) \\
&= \exp\left(\mu \sum_{j=1}^{n} s_j + \frac{1}{2}\sigma^2 \sum_{j=1}^{n} s_j^2\right) \\
&= \exp\left(\mu t_{n+1} + \frac{1}{2}\sigma^2 \frac{t_{n+1}^2}{n}\right) \exp\left(\frac{1}{2}\sigma^2 \sum_{j=1}^{n}\left(t_j - \frac{\sum_{i=1}^{n} t_i}{n}\right)^2\right) \\
&= M_{\bar{X}}(t_{n+1}) M_{X_1-\bar{X}, X_2-\bar{X}, \cdots, X_n-\bar{X}}(t_1, t_2, \cdots, t_n).
\end{aligned}$$

Here, we use the observation that

$$\begin{aligned}
M_{X_1-\bar{X}, X_2-\bar{X}, \cdots, X_n-\bar{X}}(t_1, t_2, \cdots, t_n) &= M_{X_1-\bar{X}, X_2-\bar{X}, \cdots, X_n-\bar{X}, \bar{X}}(t_1, t_2, \cdots, t_n, 0) \\
&= \exp\left(\frac{1}{2}\sigma^2 \sum_{j=1}^{n}\left(t_j - \frac{\sum_{i=1}^{n} t_i}{n}\right)^2\right)
\end{aligned}$$

and

$$M_{\bar{X}}(t_{n+1}) = M_{X_1-\bar{X}, X_2-\bar{X}, \cdots, X_n-\bar{X}, \bar{X}}(0, 0, \cdots, 0, t_{n+1}) = \exp\left(\mu t_{n+1} + \frac{1}{2}\sigma^2 \frac{t_{n+1}^2}{n}\right).$$

Therefore, $(X_1 - \bar{X}, X_2 - \bar{X}, \cdots, X_n - \bar{X})$ and $\bar{X}$ are independent. Consequently, the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and $\bar{X}$ are independent.

4. TO ADD: ADDITIONAL CONTENT

Computation of DF for constant RV and for RV with two values (immediately after the definition)