Probability

Probability is the branch of mathematics dealing with numerical description of how likely something is to happen.

In other words, probability is the measure of the likelihood that an event will occur.

Definition 1. (1) *A set E is said to be countable if there exists a bijective map from the set of natural numbers $\mathbb{N}$ to E.*
(2) *A set E is said to be uncountable if it is neither finite nor countable.*

Example 2. (1) *Define $f : \mathbb{N} \longrightarrow \mathbb{N}$ by $f(n) = n$. Clearly f is one -one and onto. Thus, $\mathbb{N}$ is countable.*
(2) *Let $\mathbb{Z}$ denotes the set of integers. Define $f : \mathbb{N} \longrightarrow \mathbb{Z}$ by*

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

Clearly f is one-one and onto. Thus, $\mathbb{Z}$ is countable.
(3) *The set of all rational numbers $\mathbb{Q}$ is also countable. Prove!*
(4) *The set of real numbers $\mathbb{R}$ as well as intervals (excluding one point set) in $\mathbb{R}$ are uncountable. Prove!*

Definition 3 (Random experiment). *A random experiment is an experiment in which*

(1) *the set of all possible outcomes of the experiment is known in advance;*
(2) *the outcome of a particular trial of the experiment cannot be predicted in advance;*
(3) *the experiment can be repeated under identical conditions.*

Definition 4 (Sample Space). *The set of all possible outcomes of a random experiment is called the sample space. We will denote the sample space of a random experiment by $\mathcal{S}$. For example:*

(1) *For tossing a fair (unbiased) coin, the sample space $\mathcal{S}$ is $\{H, T\}$, where H means that the outcome of the toss is a head and T means that it is a tail.*
(2) *For rolling a fair die, the sample space $\mathcal{S}$ is $\{1, 2, 3, 4, 5, 6\}$.*
(3) *For simultaneously flipping a coin and rolling a die, the sample space $\mathcal{S}$ is $\{H, T\} \times \{1, 2, 3, 4, 5, 6\}$.*
(4) *For flipping two coins, the sample space $\mathcal{S}$ is $\{(H, H), (H, T), (T, H), (T, T)\}$.*
(5) *For rolling two dice, the sample space $\mathcal{S}$ is $\{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$.*

Definition 5 (σ-algebra). *A non-empty collection $\mathcal{F}$ of subsets of $\mathcal{S}$ is called a σ-algebra (or σ-field) if*

(1) $\mathcal{S} \in \mathcal{F}$;
(2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;
(3) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Event and Event space: An event is a subset of the sample space $\mathcal{S}$. We say that the event E occurs when the outcome of the random experiment lies in E.

In general, any subset of $\mathcal{S}$ is not necessarily an event, rather an event is a special subset. The event space (set of all events), denoted by Σ, is a subset of the power set of $\mathcal{S}$. An event space must be a σ-algebra.

In the next remark the event space will be fixed for different sample spaces. This will be used throughout the course.

Remark 6. (1) ***If the sample space $\mathcal{S}$ is a finite or a countable set, then we will take $\Sigma = \mathcal{P}(\mathcal{S})$, where $\mathcal{P}(\mathcal{S})$ is the power set of $\mathcal{S}$.***

(2) *Let $\mathbb{B}_{\mathbb{R}}$ denote the set which contains all open intervals, closed intervals, countable unions of open intervals, countable unions of closed intervals, countable intersections of open intervals, and countable intersections of closed intervals.*

If the sample space $\mathcal{S} = \mathbb{R}$, then we will take the event space $\Sigma = \mathbb{B}_{\mathbb{R}}$

(3) *Let I be any interval. Let $\mathbb{B}_I$ denote a set which contains all open intervals contained in I, all closed intervals contained in I, all countable unions of open intervals contained in I, all countable unions of closed intervals contained in I, all countable intersections of open intervals contained in I, and all countable intersections of closed intervals contained in I.*

If the sample space $\mathcal{S} = I$, then we will take the event space $\Sigma = \mathbb{B}_I$.

For any two events E and F, the event $E \cup F$ consists of all outcomes that are either in E or in F, i.e., the event $E \cup F$ will occur if either E or F occurs. The Event $E \cap F$ consists of all outcomes which are both in E and F, i.e., the event $E \cap F$ will occur if both E and F occur.

Mutually exclusive events: Two events E_1 and E_2 are said to be mutually exclusive if they cannot occur simultaneously, i.e., if $E_1 \cap E_2 = \emptyset$.

Similarly, we can define union and intersection of more than two events.

For any event E, the event E^c (complement of E) consists of all outcomes in the sample space S that are not in E, i.e., E^c will occur if E does not occur. For example, let $E = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$, i.e., E is the event that the sum of the dice is equal to seven, then E^c will occur if the sum of the dice is not equal to seven.

Definition 7. *Consider a random experiment with sample space $\mathcal{S}$. For each event E, we assume that a real number $P(E)$ is assigned which satisfies the following three axioms:*

(1) *$0 \leq P(E) \leq 1$ or simply $P(E) \geq 0$, for all events E;*
(2) *$P(\mathcal{S}) = 1$;*
(3) *If $E_1, E_2, \ldots$ is a countably infinite collection of mutually exclusive events, that is, $E_i \cap E_j = \emptyset$ for $i \neq j$, then $P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.*

The real number $P(E)$ is known as the probability of the event E.

Remark 8. *It is clear that P is a function from the events space Σ to $[0,1]$ which satisfies the axioms (1), (2) and (3). We will call P a **probability function** and the triple $(\mathcal{S}, \Sigma, P)$ is called the probability space.*

Example 9. (1) *Let $\mathcal{S} = \{1,2,3,\ldots\}$. Define P on $\mathcal{P}(\mathcal{S})$ as follows:*

$$P(i) = \frac{1}{2^i}, i = 1, 2, \ldots.$$

Then P defines a probability (verify this!).

(2) *Let $\mathcal{S} = (0,\infty)$. Define P on $\mathbb{B}_{(0,\infty)}$ as follows: For each interval $I \subset \mathcal{S}$*

$$P(I) = \int_I e^{-x} dx.$$

Then P defines a probability (verify this!).

(3) *Let $\mathcal{S} = [0,1]$. Define P on $\mathbb{B}_{[0,1]}$ as follows: For each interval $I \subset \mathcal{S}$*

$$P(I) = \textit{length of } I.$$

Then P defines a probability (verify!).

Assigning Probabilities:

(1) Suppose $\mathcal{S}$ is a finite set containing n elements. Then it is sufficient to assign probability to each event containing single element. Thus for any events E, we have $P(E) = \sum_{w \in E} P(w)$. One such assignment is the equally likely assignment or the assignment of uniform probabilities. According to this assignment, $P(w) = \frac{1}{n}$, for every $w \in \mathcal{S}$ and $P(E) = \frac{\text{number of elements in } E}{n}$.
(2) If $\mathcal{S}$ is a countable set, one can not make an equally likely assignment of probabilities. It suffices to make the assignment for each event containing single element. Then for any event E, define $P(E) = \sum_{w \in E} P(w)$.
(3) If $\mathcal{S}$ is an uncountable set, then again one can not make an equally likely assignment of probabilities.

Theorem 10. *Let $(\mathcal{S}, \Sigma, P)$ be a probability space. Then*

(1) $P(\emptyset) = 0$;
(2) *For mutually exclusive events $E_1, E_2, \ldots, E_n$, we have* $P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$;
(3) $P(E^c) = 1 - P(E)$;
(4) *For $E_1 \subseteq E_2$, we have $P(E_1) \leq P(E_2)$ and $P(E_2 - E_1) = P(E_2) - P(E_1)$;*
(5) $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.

Proof. (1) Let $E_1 = \mathcal{S}$ and $E_i = \emptyset$, $i = 2, 3, \ldots$. Then $P(E_1) = 1, E_1 = \cup_{i=1}^{\infty} E_i$ and $E_i \cap E_j = \emptyset$ for $i \neq j$. Therefore

$$\begin{aligned} 1 = P(E_1) &= P(\cup_{i=1}^{\infty} E_i) \\ &= \sum_{i=1}^{\infty} P(E_i) \\ &= 1 + \sum_{i=2}^{\infty} P(\emptyset) \end{aligned}$$

$$\Rightarrow \sum_{i=2}^{\infty} P(\emptyset) = 0$$

This shows that the constant series $\sum_{i=2}^{\infty} P(\emptyset)$ converges to 0. Hence, $P(\emptyset) = 0$, otherwise the constant series $\sum_{i=2}^{\infty} P(\emptyset)$ can not be convergent.

(2) Let $E_i = \emptyset$, $i = n+1, n+2, \ldots$. Then $E_i \cap E_j = \emptyset$ for $i \neq j$ and $P(E_i) = 0$, $i = n+1, n+2, \ldots$. Therefore, $P(\cup_{i=1}^{n} E_i) = P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i) = \sum_{i=1}^{n} P(E_i)$ (since $P(E_i) = 0$, $i = n+1, n+2, \ldots$).

(3) Since $E \cup E^c = \mathcal{S}$ and $E \cap E^c = \emptyset$, $1 = P(E \cup E^c)$. Thus $1 = P(E) + P(E^c)$ (by using (2)). Hence $P(E^c) = 1 - P(E)$.

(4) Since $E_2 = E_1 \cup (E_2 - E_1)$ and $E_1 \cap (E_2 - E_1) = \emptyset$, $P(E_2) = P(E_1 \cup (E_2 - E_1)) = P(E_1) + P(E_2 - E_1)$. This implies that $P(E_2 - E_1) = P(E_2) - P(E_1)$.

(5) Since $E_1 \cup E_2 = E_1 \cup (E_2 - E_1)$ and $E_1 \cap (E_2 - E_1) = \emptyset$, $P(E_1 \cup E_2) = P(E_1 \cup (E_2 - E_1)) = P(E_1) + P(E_2 - E_1)$ $\ldots(i)$
Also since $E_2 = (E_1 \cap E_2) \cup (E_2 - E_1)$ and $(E_1 \cap E_2) \cap (E_2 - E_1) = \emptyset$, $P(E_2) = P(E_1 \cap E_2) + P(E_2 - E_1) \Rightarrow P(E_2 - E_1) = P(E_2) - P(E_1 \cap E_2)$ $\ldots(ii)$
Thus, by equation (i) and (ii), we have

$$\boxed{P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)}$$

□

Inclusion-exclusion identity: For events E_1, E_2 and E_3 we have

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) &= P((E_1 \cup E_2) \cup E_3) = P(E_1 \cup E_2) + P(E_3) - P((E_1 \cup E_2) \cap E_3) \\ &= P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P((E_1 \cap E_3) \cup (E_2 \cap E_3)) \\ &= P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_2 \cap E_3) \\ &\quad + P(E_1 \cap E_2 \cap E_3) \end{aligned}$$

Inductively, for any n events $E_1, E_2, \ldots, E_n$, we have

$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k}^{n} P(E_i \cap E_j \cap E_k) - \cdots + (-1)^{n+1} P(E_1 \cap E_2 \cap \cdots \cap E_n).$$

This identity is known as the inclusion-exclusion identity.

Exhaustive events: The countable collection $\{E_i \mid i \in \wedge\}$ of events is said to be exhaustive if $P(\cup_{i \in \wedge} E_i) = 1$, where $\wedge$ is an index set.

Definition 11. *Let $(\mathcal{S}, \Sigma, P)$ be a probability space and (E_n) be a sequence of events in Σ.*

(1) *We say that sequence (E_n) is increasing (written as $E_n \uparrow$) if $E_n \subseteq E_{n+1}$, $n = 1, 2, \ldots$;*
(2) *We say that sequence (E_n) is decreasing (written as $E_n \downarrow$) if $E_{n+1} \subseteq E_n$, $n = 1, 2, \ldots$;*
(3) *We say that the sequence (E_n) is monotone if either $E_n \uparrow$ or $E_n \downarrow$.*
(4) *If $E_n \uparrow$, we define $Lim_{n\to\infty} E_n = \cup_{n=1}^{\infty} E_n$*
(5) *If $E_n \downarrow$, we define $Lim_{n\to\infty} E_n = \cap_{n=1}^{\infty} E_n$*

Theorem 12. *(**Continuity of Probability**) Let (A_n) be a sequence of monotone events. Then*

$$P(Lim_{n\to\infty} E_n) = \lim_{n\to\infty} P(E_n),$$

where $\lim_{n\to\infty} P(E_n)$ denotes the limit of the real sequence $(P(E_n))$.

Proof. Exercise □

Conditional Probability and Bayes Formula

Consider a random experiment with sample space $\mathcal{S}$. In many situations we may not be interested in the whole sample space but a subset of the sample space.

For example, suppose that we toss two fair dice. Then $\mathcal{S} = \{(i,j) : i,j \in \{1,2,3,4,5,6\}\}$ and $P((i,j)) = \frac{1}{36}$. Suppose we observe that the first die is four. If the first die is a four, then there can be at most six possible outcomes, namely, $(4,1)$, $(4,2)$, $(4,3)$, $(4,4)$, $(4,5)$, $(4,6)$. So given that the first die is a four, the probability that the sum of the two dice equals six is $1/6$. Let E and F be the event that the sum of the dice is six and the event that the first die is a four respectively, then the probability obtained is called the conditional probability that E occurs given that F has occurred.

Definition 1. *Let $(\mathcal{S}, \Sigma, P)$ be a probability space and F be the fixed event with $P(F) > 0$. Then the probability that an event E occurs given that F has occurred is called the conditional probability of event E given that the outcomes of the experiment is in F or simply the conditional probability of E given F. It is denoted by $P(E|F)$ and is defined by*

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

Theorem 2. *Let $(\mathcal{S}, \Sigma, P)$ be a probability space and F be the fixed event with $P(F) > 0$. Then $(\mathcal{S}, \Sigma, P_F)$, where $P_F(E) = P(E|F)$ for all $E \in \Sigma$, is a probability space.*

Proof. Exercise □

Example 3. *Suppose cards numbered one to ten are placed in a hat, mixed up, and then one of the card is drawn. If we are told that the number on the drawn card is at least five, then what is the conditional probability that it is ten?*

Solution: Let E be the event that the number on the drawn card is ten and F be the event that the number on the drawn card is at least five.

Then $P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{1/10}{6/10} = 1/6$.

Example 4. *Suppose that each of three men at a party throws his hat into the center of the room. The hats are first mixed up and then each man randomly select a hat. What is the probability that none of the three men select his own hat?*

Solution: For $i = 1,2,3$, let E_i be the event that the i-th man selects his own hat. Then $P(E_i) = 1/3$, for each $i = 1,2,3$.

Given that the i-th man has selected his own hat, then there remain two hats that the j-th man may select, and as one of these two is his own hat, so $P(E_j|E_i) = 1/2$ for $i \neq j$. Therefore, $P(E_i \cap E_j) = P(E_i)P(E_j|E_i) = 1/6$ for $i \neq j$. Also $P(E_1 \cap E_2 \cap E_3) = P(E_1 \cap E_2)P(E_3|E_1 \cap E_2) = \frac{1}{6}P(E_3|E_1 \cap E_2)$.

However, given that the first two men get their own hats, it follows that the third man must also get his own hat (since there are no other hats left). That is, $P(E_3|E_1 \cap E_2) = 1$. So $P(E_1 \cap E_2 \cap E_3) = \frac{1}{6}$

Now, the probability that at least one of them selects his own hat is, $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_2 \cap E_3) - P(E_1 \cap E_3) + P(E_1 \cap E_2 \cap E_3) = 2/3$.

Hence the probability that none of the three men select his own hat is $1 - P(E_1 \cup E_2 \cup E_3) = 1/3$.

Example 5. *Suppose an urn contains seven black balls and five white balls. We draw two balls from the urn without replacement. Assuming that each ball in the urn is equally likely to be drawn, what is the probability that both drawn balls are black?*

Solution: Let E be the event that the first drawn ball is black and F be the event that the second drawn ball is black.

Then $P(E) = 7/12$ and $P(F|E) = 6/11$. Therefore, $P(E \cap F) = P(E)P(F|E) = 42/132$.

Independent Events: Two events E and F are said to be independent if $P(E \cap F) = P(E)P(F)$. Two events E and F are said to be dependent if E and F are not independent.

Remark 6. (1) *Suppose $P(F) > 0$. Then $P(E|F) = \frac{P(E \cap F)}{P(F)}$. Assume E and F are independent i.e., $P(E \cap F) = P(E)P(F)$. Then $P(E|F) = P(E)$. This implies that if $P(F) > 0$, then E and F are independent if and only if $P(E|F) = P(E)$.*

In other words, E and F are independent if and only if the availability of the information that event F has occurred does not alter the probability of occurrence of event E.

(2) *If $P(F) = 0$, then $P(E \cap F) = 0 = P(E)P(F)$, for every event E, that is, if $P(F) = 0$, then any event E and F are independent.*

Definition 7. (1) *Events $\{E_i : i \in \wedge\}$, where $\wedge$ is an index set, are said to be pairwise independent if any pair of events E_i and E_j, $i \neq j$, are independent, i.e., $P(E_i \cap E_j) = P(E_i)P(E_j)$ $i \neq j$.*

(2) *Events $E_1, E_2, \ldots, E_n$ are said to be independent if for any sub-collection $\{E_{i_1}, E_{i_2}, \ldots, E_{i_k}\}$, $2 \leq k \leq n$, we have $P(\cap_{j=1}^{k} E_{i_j}) = \prod_{j=1}^{k} P(E_{i_j})$.*

By definition, it is clear that independence of a finite collection of events implies pairwise independence of events but the converse need not be true. Also to verify that n events $E_1, E_2, \cdots, E_n$ are independent, we have to verify $2^n - n - 1$ conditions.

The following example shows that pairwise independence does not imply independence.

Example 8. *Let a sample space $\mathcal{S} = \{1, 2, 3, 4\}$ with $P(i) = 1/4$, for each $i = 1, 2, 3, 4$. Consider the following events: $A = \{1, 4\}$, $B = \{2, 4\}$ and $C = \{3, 4\}$. Then $P(A) = P(B) = P(C) = 1/2$, $P(A \cap B) = P(B \cap C) = P(C \cap A) = 1/4$, and $P(A \cap B \cap C) = 1/4$. Clearly A, B and C are pairwise independent but not independent.*

Example 9. *Suppose we toss two fair dice. Let E_1 be the event that the sum of the dice is six, E_2 be the event that the sum of the dice is seven, and F be the event that the first die equals four. Then*

$$P(E_1) = 5/36, P(E_2) = 1/6, P(F) = 1/6 \text{ and } P(E_1 \cap F) = P(E_2 \cap F) = 1/36.$$

Since $P(E_1 \cap F) \neq P(E_1)P(F)$, E_1 and F are not independent. On the other hand, since $P(E_2 \cap F) = P(E_2)P(F)$, E_2 and F are independent.

Intuitively it is clear, because chance of getting a total of six depends on the outcome of the first die and hence E_1 and F cannot be independent while for every outcome of the first die, we have a chance of getting a total of seven and hence E_2 and F are independent.

Bayes Formula

Let E and F be two events. Then we have

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

Since $E \cap F$ and $E \cap F^c$ are mutually exclusive,

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

Suppose $0 < P(F) < 1$. Then

$$P(E) = P(E|F)P(F) + P(E|F^c)P(F^c)$$

Now we will generalize this as follows:

Suppose that $F_1, F_2, \cdots, F_n$ are mutually exclusive and exhaustive events, that is, $F_i \cap F_j = \emptyset$ for $i \neq j$ and $P(\cup_{i=1}^n F_i) = 1$.

Let E be any event and $F = \cup_{i=1}^n F_i$. Since $P(F) = 1$, $P(F^c) = 0$. Then

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

Suppose $P(F_i) > 0$, for all $1 \leq i \leq n$. Then we have

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

This formula is called the **total probability rule**.

Now, for $P(E) > 0$, we have

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

$$\boxed{P(F_j|E) = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^n P(E|F_i)P(F_i)}}$$

This is called the **Bayes formula**.

Example 10. *Consider two urns. The first contains two white and seven black ball, and the second contains five white and six black balls. We flip a fair coin and then draw ball from the first urn or the second urn depending on whether the outcome was heads or tails. What is the conditional probability that the outcome of the toss was heads given that a white ball was selected?*

Solution: Let W be the event that a white ball is drawn and H be the event that the coin comes up head.

Then $P(H|W) = \frac{P(H \cap W)}{P(W)} = \frac{P(W|H)P(H)}{P(W)} = \frac{P(W|H)P(H)}{P(W|H)P(H)+P(W|H^c)P(H^c)}$

Hence $P(H|W) = \frac{2/9 \times 1/2}{(2/9 \times 1/2)+(5/11 \times 1/2)} = \frac{22}{67}$.

Example 11. *Urn 1 contains one white and two black marbles, Urn 2 contains one black and two white marbles, and Urn 3 contains three black and three white marbles. A die is rolled. If a* $1, 2$ *or* 3 *shows up, Urn 1 is selected, if a* 4 *shows up, Urn 2 is selected, and if a* 5 *or* 6 *shows up, Urn 3 is selected. A marble is then drawn at random from the urn selected. Let* A *be the event that the drawn marble is white and* U, V, W *respectively denotes the events that the urn selected is* $1, 2, 3$*. What is the probability that urn 2 is selected given that the marble drawn is white?*

Solution: $P(V|A) = \frac{P(A\cap V)}{P(A)} = \frac{P(A|V)P(V)}{P(A|U)P(U)+P(A|V)P(V)+P(A|W)P(W)}$

Hence $P(V|A) = \frac{1/6\times 2/3}{(3/6\times 1/3)+(1/6\times 2/3)+(2/6\times 3/6)} = \frac{1}{4}$.

Example 12. *Urn* U_1 *contains* 4 *white and* 6 *black balls, and Urn* U_2 *contains* 6 *white and* 4 *black balls. A fair die is cast and Urn* U_1 *is selected if the upper face of die shows five or six dots otherwise Urn* U_2 *is selected. A ball is drawn at random from the selected urn.*

(1) *Given that the drawn ball is white, find the conditional probability that it came from Urn* U_1.
(2) *Given that the drawn ball is white, find the conditional probability that it came from Urn* U_2.

Solution: Let W be the event that the drawn ball is white, E_1 be the event that the Urn U_1 is selected and E_2 be the event that the Urn U_2 is selected. Clearly E_1 and E_2 are mutually exclusive and exhaustive events.

(1) $P(E_1|W) = \frac{P(E_1\cap W)}{P(W)}) = \frac{P(W|E_1)P(E_1)}{P(W|E_1)P(E_1)+P(W|E_2)P(E_2)} = \frac{4/10\times 2/6}{(4/10\times 2/6)+(6/10\times 4/6)} = \frac{1}{4}$
(2) $P(E_2|W) = \frac{P(E_2\cap W)}{P(W)}) = \frac{P(W|E_2)P(E_2)}{P(W|E_1)P(E_1)+P(W|E_2)P(E_2)} = \frac{6/10\times 4/6}{(4/10\times 2/6)+(6/10\times 4/6)} = \frac{3}{4}$

Random Variable

Let $(\mathcal{S}, \Sigma, P)$ be a probability space. Since P is a set function, it is not very easy to handle. Also in many situations, one may not be interested in the sample space rather one may be interested in some numerical characteristics of the sample space. For example, when a coin is tossed n-times, which replication resulted in heads is not of much interest. Rather, one is interested in the number of heads, and consequently, the number of tails, that appear out of n tosses.

It is therefore desirable to introduce a point function on the sample space so that we can use the theory of calculus or real analysis to study the properties of P.

Definition 1. *A function $X : \mathcal{S} \longrightarrow \mathbb{R}$ is called a random variable (RV) if $X^{-1}(B) \in \Sigma$, for all $B \in \mathbb{B}_{\mathbb{R}}$, that is, $X^{-1}(B) = \{w \in \mathcal{S} : X(w) \in B\}$ is an event.*

Notations.

We will use the following notations throughout the course.

- For $B \in \mathbb{B}_{\mathbb{R}}$, $\{X \in B\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : X(w) \in B\} \stackrel{\text{def}}{=} X^{-1}(B)$;
- $\{a < X \leq b\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : a < X(w) \leq b\} \stackrel{\text{def}}{=} X^{-1}((a, b])$;
- $\{a \leq X \leq b\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : a \leq X(w) \leq b\} \stackrel{\text{def}}{=} X^{-1}([a, b])$;
- $\{a < X < b\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : a < X(w) < b\} \stackrel{\text{def}}{=} X^{-1}((a, b))$;
- $\{a \leq X < b\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : a \leq X(w) < b\} \stackrel{\text{def}}{=} X^{-1}([a, b))$;
- $\{X = a\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : X(w) = a\} \stackrel{\text{def}}{=} X^{-1}(\{a\})$;
- $\{X \leq a\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : X(w) \leq a\} \stackrel{\text{def}}{=} X^{-1}((-\infty, a])$;
- $\{X < a\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : X(w) < a\} \stackrel{\text{def}}{=} X^{-1}((-\infty, a))$;
- $\{X \geq a\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : X(w) \geq a\} \stackrel{\text{def}}{=} X^{-1}([a, \infty))$;
- $\{X > a\} \stackrel{\text{def}}{=} \{w \in \mathcal{S} : X(w) > a\} \stackrel{\text{def}}{=} X^{-1}((a, \infty))$.

Remark 2. (1) *X is a random variable if and only if for each $x \in \mathbb{R}$, $\{X \leq x\} \in \Sigma$.*

(2) *If $\Sigma = \mathcal{P}(\mathcal{S})$, then any function $X : \mathcal{S} \longrightarrow \mathbb{R}$ is a random variable.*

(3) *Let $(\mathcal{S}, \Sigma, P)$ be a probability space and $X : \mathcal{S} \longrightarrow \mathbb{R}$ be a random variable. Then the random variable X induces a probability space $(\mathbb{R}, \mathbb{B}_{\mathbb{R}}, P_X)$, where $P_X(B) = P(\{w \in \mathcal{S} : X(w) \in B\})$, for all $B \in \mathbb{B}_{\mathbb{R}}$.*

Example 3. *Suppose that a fair coin is independently flipped thrice. Then $\mathcal{S} = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$ and $P(E) = \frac{\text{number of elements in } E}{8}$, for every $E \in \mathcal{P}(\mathcal{S})$. Define $X : \mathcal{S} \longrightarrow \mathbb{R}$ by $X(w) =$ number of heads, i.e.,*

$$X(w) = \begin{cases} 0, & w = \{TTT\} \\ 1, & w \in \{HTT, TTH, THT\} \\ 2, & w \in \{HHT, THH, HTH\} \\ 3, & w = \{HHH\}. \end{cases}$$

Clearly X is a random variable. The induced probability space is $(\mathbb{R}, \mathbb{B}_{\mathbb{R}}, P_X)$, where $P_X(\{0\}) = P_X(\{3\}) = \frac{1}{8}$, $P_X(\{1\}) = P_X(\{2\}) = \frac{3}{8}$, and $P_X(B) = \sum_{i \in \{0,1,2,3\} \cap B} P_X(\{i\})$, for all $B \in \mathbb{B}_{\mathbb{R}}$.

Definition 4. *Let $(\mathcal{S}, \Sigma, P)$ be a probability space and $X : \mathcal{S} \longrightarrow \mathbb{R}$ be a random variable. The function $F_X : \mathbb{R} \longrightarrow \mathbb{R}$, defined by,*

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

*is called the **cumulative distribution function** (c.d.f) or the **distribution function** (d.f) of the random variable X.*

Theorem 5. *Let F_X be the cumulative distribution function of a random variable X. Then*

(1) *F_X is non-decreasing;*
(2) *F_X is right continuous;*
(3) *$F_X(-\infty) \stackrel{\text{def}}{=} \lim_{x \to -\infty} F_X(x) = 0$ and $F_X(\infty) \stackrel{\text{def}}{=} \lim_{x \to \infty} F_X(x) = 1$.*

Proof. (1) Let $x_1 < x_2$. Then $(-\infty, x_1] \subset (-\infty, x_2]$. Then by the properties of the probability function, we have

$$F_X(x_1) = P(\{X \leq x_1\}) \leq P(\{X \leq x_2\}) = F_X(x_2).$$

(2) Fix $a \in \mathbb{R}$. Since F_X is non-decreasing, $F_X(a+) = \lim_{x \to a+} F_X(x)$ exists. Therefore

$$F_X(a+) = \lim_{n \to \infty} F_X(a + \frac{1}{n}) = \lim_{n \to \infty} P(\{X \leq a + \frac{1}{n}\}).$$

Let $E_n = \{w \in \mathcal{S} : X(w) \in (-\infty, a + \frac{1}{n}]\}$. Then E_n is an decreasing sequence of events and $\text{Lim}_{n \to \infty} E_n = \cap_{n=1}^{\infty} E_n = \{w \in \mathcal{S} : X(w) \in (-\infty, a]\}$. Now by using continuity of probability, we have

$$\begin{aligned} F_X(a+) &= \lim_{n \to \infty} P(\{X \leq a + \frac{1}{n}\}) \\ &= \lim_{n \to \infty} P(E_n) \\ &= P(\text{Lim}_{n \to \infty} E_n) \\ &= P(\{X \in (-\infty, a]\}) \\ &= P(\{X \leq a\}) \\ &= F_X(a) \end{aligned}$$

(3) Let $A_n = \{w \in \mathcal{S} : X(w) \in (-\infty, -n]\}$ and $B_n = \{w \in \mathcal{S} : X(w) \in (-\infty, n]\}$. Then A_n and B_n are decreasing and increasing sequence of events, respectively. Also $\text{Lim}_{n \to \infty} A_n = \cap_{n=1}^{\infty} A_n = \emptyset\}$ and $\text{Lim}_{n \to \infty} B_n = \cup_{n=1}^{\infty} B_n = \{w \in \mathcal{S} : X(w) \in \mathbb{R}\} = \mathcal{S}$. Therefore, by using continuity of probability, we have

$$\begin{aligned} F_X(-\infty) &= \lim_{n \to \infty} F_X(-n) \\ &= \lim_{n \to \infty} P(\{X \in (-\infty, -n]\}) \\ &= \lim_{n \to \infty} P(A_n) \\ &= P(\text{Lim}_{n \to \infty} A_n) \\ &= P(\emptyset) = 0, \end{aligned}$$

and

$$\begin{aligned} F_X(\infty) &= \lim_{n\to\infty} F_X(n) \\ &= \lim_{n\to\infty} P(\{X \in (-\infty, n]\}) \\ &= \lim_{n\to\infty} P(B_n) \\ &= P(\mathrm{Lim}_{n\to\infty} B_n) \\ &= P(\mathcal{S}) = 1. \end{aligned}$$

□

Remark 6. (1) *Let* $E_n = \{w \in \mathcal{S} : X(w) \in (-\infty, a - \frac{1}{n}]\} = \{X \leq a - \frac{1}{n}\}$. *Then* E_n *is an increasing sequence of events and* $Lim_{n\to\infty} E_n = \cup_{n=1}^{\infty} E_n = \{w \in \mathcal{S} : X(w) \in (-\infty, a)\} = \{X < a\}$. *Now by using continuity of probability, we have*

$$\begin{aligned} P(\{X < a\}) &= P(Lim_{n\to\infty} E_n) \\ &= \lim_{n\to\infty} P(E_n) \\ &= \lim_{n\to\infty} P(\{X \leq a - \frac{1}{n}\}) \\ &= \lim_{n\to\infty} F_X(a - \frac{1}{n}) \\ &= F_X(a-). \end{aligned}$$

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

(2) *For* $-\infty < a < b < \infty$, *we have*

(a) $P(\{a < X \leq b\}) = P(\{X \in ((-\infty, b] - (-\infty, a])\}) = P(\{X \leq b\}) - P(\{X \leq a\}) = F_X(b) - F_X(a)$.

(b) $P(\{a < X < b\}) = P(\{X \in ((-\infty, b) - (-\infty, a])\}) = P(\{X < b\}) - P(\{X \leq a\}) = F_X(b-) - F_X(a)$.

(c) $P(\{a \leq X < b\}) = P(\{X \in ((-\infty, b) - (-\infty, a))\}) = P(\{X < b\}) - P(\{X < a\}) = F_X(b-) - F_X(a-)$.

(d) $P(\{a \leq X \leq b\}) = P(\{X \in ((-\infty, b] - (-\infty, a))\}) = P(\{X \leq b\}) - P(\{X < a\}) = F_X(b) - F_X(a-)$.

(3) *For* $-\infty < a < \infty$, *we have*

(a) $P(\{X \geq a\}) = P(\{X \in (\mathbb{R} - (-\infty, a))\}) = P(\{X \in \mathbb{R}\}) - P(\{X < a\}) = 1 - F_X(a-)$.

(b) $P(\{X > a\}) = P(\{X \in (\mathbb{R} - (-\infty, a]))\}) = P(\{X \in \mathbb{R}\}) - P(\{X \leq a\}) = 1 - F_X(a)$.

(4) *The distribution function* F_X *has atmost countable number of discontinuities.*

Example 7. *Let* $(\mathcal{S}, \Sigma, P)$ *be a probability space. Define* $X : \mathcal{S} \longrightarrow \mathbb{R}$ *by* $X(w) = c$, *for all* $x \in \mathcal{S}$, *where* c *is a fixed real number. Clearly,* X *is a random variable and the cumulative distribution function of* X *is*

$$F_X(x) = P(\{X \leq x\}) = \begin{cases} 0, & x < c \\ 1, & x \geq c. \end{cases}$$

Example 8. *Let $(\mathcal{S}, \Sigma, P)$ be a probability space and E be an event. Define $I_E : \mathcal{S} \longrightarrow \mathbb{R}$ by $I_E(w) = 1$, if $w \in E$ and $I_E(w) = 0$, if $w \notin E$. The function I_E is called the indicator function or characteristic function of E and is sometimes denoted by 1_E or χ_E. We have*

$$\{I_E \leq a\} = I_E^{-1}((-\infty, a]) = \begin{cases} \emptyset, & a < 0 \\ E^c, & 0 \leq a < 1 \\ \mathcal{S}, & a \geq 1. \end{cases}$$

Clearly, I_E is a random variable and the cumulative distribution function of I_E is

$$F_{I_E}(a) = P(\{I_E \leq a\}) = \begin{cases} 0, & a < 0 \\ P(E^c), & 0 \leq a < 1 \\ 1, & a \geq 1. \end{cases}$$

Example 9. *Let $\mathcal{S} = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$ with $P(E) = \frac{\text{number of elements in } E}{8}$, for every $E \in \mathcal{P}(\mathcal{S})$. Let $X : \mathcal{S} \longrightarrow \mathbb{R}$ be a random variable, defined by $X(w) =$ number of heads. Then the cumulative distribution function of X is*

$$F_X(x) = P(\{X \leq x\}) = \sum_{i \in \{0,1,2,3\} \cap (-\infty, x]} P(\{X = i\}) = \begin{cases} 0, & x < 0 \\ \frac{1}{8}, & 0 \leq x < 1 \\ \frac{1}{2}, & 1 \leq x < 2 \\ \frac{7}{8}, & 2 \leq x < 3 \\ 1, & x \geq 3. \end{cases}$$

Example 10. *Consider the probability space $(\mathbb{R}, \mathbb{B}_{\mathbb{R}}, P)$ with $P(B) = \int_0^\infty e^{-t} I_B(t) dt$, where I_B is the indicator function of B. Define $X : \mathbb{R} \longrightarrow \mathbb{R}$ by*

$$X(w) = \begin{cases} 0, & w \leq 0 \\ \sqrt{w}, & w > 0. \end{cases}$$

We have

$$\{X \leq x\} = X^{-1}((-\infty, x]) = \begin{cases} \emptyset, & x < 0 \\ (-\infty, x^2], & x \geq 0. \end{cases}$$

Thus X is a random variable. Now, the cumulative distribution function of X is

$$F_X(x) = P(\{X \leq x\}) = \begin{cases} P(\emptyset), & x < 0 \\ P((-\infty, x^2]), & x \geq 0. \end{cases}$$

Thus

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \int_0^{x^2} e^{-t} dt, & x \geq 0 \end{cases} = \begin{cases} 0, & x < 0 \\ 1 - e^{-x^2}, & x \geq 0. \end{cases}$$

Definition 11. *A real-valued function $F : \mathbb{R} \longrightarrow \mathbb{R}$ that is increasing, right continuous and satisfies*

$$F(-\infty) = 0 \text{ and } F(\infty) = 1$$

is called a distribution function.

Theorem 12. *Every distribution function is the distribution function of a random variable on some probability space.*

Types of Random Variables: Discrete and Continuous

Let $(\mathcal{S}, \Sigma, P)$ be a probability space with a random variable $X : \mathcal{S} \longrightarrow \mathbb{R}$, and let $(\mathbb{R}, \mathbb{B}_{\mathbb{R}}, P_X)$ be the probability space induced by X. Let F_X be the distribution function of X. It is known that F_X uniquely determine P_X and vice-versa. Thus, to study the induced probability space $(\mathbb{R}, \mathbb{B}_{\mathbb{R}}, P_X)$, it is sufficient to study the d.f. F_X.

In this course we will restrict ourselves to two types of random variables: discrete and continuous. In the first case, the RV assumes at most a countable number of values and hence its d.f is a step function. In the later case, the d.f. F_X is continuous (we will see the definition later).

Definition 1. *A random variable X is said to be of discrete type, or simply discrete, if there exists a finite or a countable set $E_X \subset \mathbb{R}$ such that $P(\{X = x\}) > 0$, $\forall\ x \in E_X$ and $P(\{X \in E_X\}) = 1$. The set E_X is called the support of the RV X.*

Remark 2. (1) *If X is any RV with the d.f. F_X, then $P(\{X = x\}) = F_X(x) - F_X(x-)$ for every $x \in \mathbb{R}$. (Prove!)*

(2) *From previous lecture, we know that F_X is discontinuous at $x \in \mathbb{R}$ if and only if $F_X(x-) < F_X(x+) = F_X(x)$. Hence, F_X has only jump discontinuities and the size of the jump at any point x of discontinuity is $P(\{X = x\}) = F_X(x) - F_X(x-)$.*

Remark 3. (1) *If X is a discrete type RV with support E_X, then*

$$P(\{X \in E_X\}) = \sum_{x \in E_X} P(\{X = x\}) = \sum_{x \in E_X} (F_X(x) - F_X(x-)) = 1.$$

(2) *The d.f. F_X is continuous at every point of E_X^c.*

Definition 4. *Let X be a discrete type random variable with support E_X. The function $f_X : \mathbb{R} \to \mathbb{R}$ defined by,*

$$f_X(x) = \begin{cases} P(\{X = x\}), \text{if } x \in E_X, \\ 0, \text{otherwise} \end{cases}$$

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

Remark 5. *Let X be a discrete type RV with support E_X, the d.f. F_X and the p.m.f. f_X.*

(1) $f_X(x) > 0$, $\forall\ x \in E_X$ *and* $f_X(x) = 0$, $\forall\ x \notin E_X$.
(2) $\sum_{x \in E_X} f_X(x) = 1$.
(3) *For $A \in \mathbb{B}_{\mathbb{R}}$, we have*

$$\begin{aligned} P_X(A) &= P_X(A \cap E_X) + P_X(A \cap E_X^c) \\ &= P_X(A \cap E_X) \\ &= \sum_{x \in A \cap E_X} f_X(x). \end{aligned}$$

(4) *For $x \in \mathbb{R}$, we have*

$$F_X(x) = \sum_{y \in (-\infty, x] \cap E_X} f_X(y).$$

Example 6. *Consider the the random variable defined as $X(w) = c$ for all $w \in \mathcal{S}$, where c is a fixed real number. Then $P(\{X = c\}) = 1$ and $E_X = \{c\}$. Hence, X is of discrete*

type. Its p.m.f. is given by

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

Example 7. *Let X be the indicator function of E, where E is an event. Then $E_X = \{0, 1\}$ and $P(\{X \in E_X\}) = 1$. Thus, X is discrete and its p.m.f. is given by*

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

Example 8. *Consider a coin that, in any flip, ends up in head with probability $\frac{1}{4}$ and in tail with probability $\frac{3}{4}$. The coin is tossed repeatedly and independently until a total of two heads have been observed. Let X denote the number of flips required to achieve this. Then $P(\{X = x\}) = 0$, if $x \notin \{2, 3, 4, \ldots\}$. For $n \in \{2, 3, 4, \ldots\}$, let $S_n = \{(w_1, w_2, \ldots, w_n) : w_n = H, w_i = H$ for one i between 1 and $n-1$, and $w_j = T$, for $j \neq i\}$. Now,*

$$\begin{aligned} P(\{X = n\}) &= \sum_{(w_1, w_2, \ldots, w_n) \in S_n} P(\{(w_1, w_2, \ldots, w_n)\}) \\ &= P(\{(w_1, w_2, \ldots, w_n)\})|S_n| \\ &= P(\{w_1\})P(\{w_2\}) \cdots P(\{w_n\})|S_n| \quad (\textit{since all events are independent}) \\ &= \frac{1}{4}\left(\frac{3}{4}\right)^{n-2} \frac{1}{4}\binom{n-1}{1} \\ &= \frac{n-1}{16}\left(\frac{3}{4}\right)^{n-2}. \end{aligned}$$

Also, $\sum_{n=2}^{\infty} P(\{X = n\}) = 1$. Thus, X is of discrete type with support $E_X = \{2, 3, 4, \ldots\}$ and p.m.f.

$$f_X(x) = \begin{cases} \frac{x-1}{16}\left(\frac{3}{4}\right)^{x-2}, \text{ if } x \in \{2, 3, 4, \ldots\}, \\ 0, \text{ otherwise.} \end{cases}$$

The d.f. of X is

$$\begin{aligned} F_X(x) &= P(\{X \leq x\}) \\ &= \begin{cases} 0, \text{ if } x < 2, \\ \frac{1}{16}\sum_{j=2}^{i}(j-1)(\frac{3}{4})^{j-2}, \text{ if } i \leq x < i+1, i = 2, 3, 4, \ldots, \end{cases} \\ &= \begin{cases} 0, \text{ if } x < 2, \\ 1 - \frac{i+3}{4}(\frac{3}{4})^{i-2}, \text{ if } i \leq x < i+1, i = 2, 3, 4, \ldots. \end{cases} \end{aligned}$$

Definition 9. *A random variable X with the d.f. F_X is said to be of continuous type, or simply continuous, if there exists an integrable function $f_X : \mathbb{R} \to \mathbb{R}$ such that $f_X(x) \geq 0$ for every $x \in \mathbb{R}$ and*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \ x \in \mathbb{R}.$$

The function f_X is called the probability density function (p.d.f.) of random variable X and the set $E_X = \{x \in \mathbb{R} : f_X(x) > 0\}$ is called the support of random variable X (or of p.d.f. f_X.

Remark 10. *Let X be a continuous type RV with the support E_X, the d.f. F_X and a p.d.f. f_X.*

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

(2) *F_X is continuous on $\mathbb{R}$. (Prove!) Therefore, $P(\{X = x\}) = 0 \; \forall \; x \in \mathbb{R}$. In general, for any countable set C, $P(\{X \in X\}) = 0$.*

(3) *Let $a, b \in \mathbb{R}$ with $a < b$. Then*

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

In general, for any $B \in \mathbb{B}_{\mathbb{R}}$, we have $P(\{X \in B\}) = \int_{-\infty}^{\infty} f_X(t) I_B(t) dt$, where I_B is the indicator function of B.

Remark 11. (1) *Suppose that the d.f. F_X of an RV X is differentiable at every $x \in \mathbb{R}$. Then*

$$F_X(x) = \int_{-\infty}^{x} F_X'(t)dt, \; x \in \mathbb{R}.$$

This implies that X is of continuous type and we may take its p.d.f to be $f_X(x) = F_X'(x)$, $x \in \mathbb{R}$.

(2) *Suppose that the d.f. F_X of an RV X is differentiable everywhere except on a countable set C. Further suppose that*

$$\int_{-\infty}^{\infty} F_X'(t) I_{C^c} dt = 1.$$

This again will imply that X is of continuous type with p.d.f. (Verify!)

$$f_X(x) = \begin{cases} F_X'(x), \text{if } x \notin C, \\ a_x, \text{if } x \in C, \end{cases}$$

where a_x, $x \in C$ are arbitrary non negative constants.

(3) *From the previous remark, it is clear that p.d.f. of a continuous random variable need not be unique.*

(4) *There are random variables that are neither of discrete type nor of continuous type. Find some examples.*

Example 12. *Let X be an RV having d.f.*

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

We observe that F_X is differentiable everywhere except at $x = 0$. Let $C = \{0\}$. Moreover,

$$\int_{-\infty}^{\infty} F_X'(t) I_{C^c} dt = \int_0^{\infty} e^{-t} = 1.$$

Hence, X is of continuous type and its p.d.f. is

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

Function of Random Variables

Let $(\mathcal{S}, \Sigma, P)$ be a probability space with a random variable $X : \mathcal{S} \longrightarrow \mathbb{R}$, and let $h : \mathbb{R} \to \mathbb{R}$ be a function. Consider the function $Z : \mathcal{S} \to \mathbb{R}$ defined by $Z(w) = h(X(w))$. It will be interesting to know when Z is an RV and what are the probabilistic properties of Z.

Remark 1. (1) *Let X be a random variable and let $h : \mathbb{R} \to \mathbb{R}$ be a function such that $h^{-1}(A) \in \mathbb{B}_{\mathbb{R}}$ for every $A \in \mathbb{B}_{\mathbb{R}}$. Then the function $Z : \mathcal{S} \to \mathbb{R}$ defined by $Z(w) = h(X(w))$ is a random variable. Prove! The function Z (written as $h \circ X$ or $h(X)$) is called a function of random variable X.*

(2) *If X is an RV, and $h : \mathbb{R} \to \mathbb{R}$ is a continuous function, then $h \circ X$ is an RV. In particular, X^2, $|X|$, $\max\{X, 0\}$ and $\sin X$ are random variables.*

(3) *If X is an RV and h is strictly monotone then $h(X)$ is an RV.*

The next two theorems deal with probability distribution of a function of random variables.

Theorem 2. *Let X be an RV of discrete type with support E_X and p.m.f. f_X. Let $h : \mathbb{R} \to \mathbb{R}$ be a function such that $h^{-1}(A) \in \mathbb{B}_{\mathbb{R}}$ for every $A \in \mathbb{B}_{\mathbb{R}}$ and let $Z : \mathcal{S} \to \mathbb{R}$ be a function defined by $Z(w) = h(X(w))$. Then Z is an RV of discrete type with support $E_Z = \{h(x) : x \in E_X\}$ and p.m.f.*

$$f_Z(z) = \begin{cases} \sum_{x \in A_z} f_X(x), \text{ if } z \in E_Z, \\ 0, \text{ otherwise} \end{cases}$$
$$= \begin{cases} P(\{X \in A_z\}), \text{ if } z \in E_Z, \\ 0, \text{ otherwise,} \end{cases}$$

where $A_z = \{x \in E_X : h(x) = z\}$.

Problem 3. *Let X be a random variable with p.m.f.*

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

Show that $Z = X^2$ is a random variable and find its p.m.f.

Solution. Clearly $E_X = \{-2, -1, 0, 1, 2, 3\}$ and $E_Z = \{0, 1, 4, 9\}$. Also,

$P(\{Z = 0\}) = P(\{X^2 = 0\}) = P(\{X = 0\}) = \frac{1}{7}$,

$P(\{Z = 1\}) = P(\{X^2 = 1\}) = P(\{X \in \{-1, 1\}\}) = \frac{1}{7} + \frac{1}{7} = \frac{2}{7}$,

$P(\{Z = 4\}) = P(\{X^2 = 4\}) = P(\{X \in \{-2, 2\}\}) = \frac{1}{7} + \frac{3}{14} = \frac{5}{14}$,

$P(\{Z = 9\}) = P(\{X^2 = 9\}) = P(\{X \in \{-3, 3\}\}) = 0 + \frac{3}{14} = \frac{3}{14}$.

Clearly Z is of discrete type.

The p.m.f. of Z is

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

Theorem 4. *Let X be a random variable of continuous type with p.d.f. f_X and support E_X such that E_X is a finite union of disjoint open intervals in $\mathbb{R}$. Let $h : \mathbb{R} \to \mathbb{R}$ function such that $h^{-1}(A) \in \mathbb{B}_{\mathbb{R}}$ for every $A \in \mathbb{B}_{\mathbb{R}}$, and h is differentiable and strictly monotone on E_X. Let $E_T = \{h(x) : x \in E_X\}$. Then $T = h(X)$ is an RV of continuous type with p.d.f.*

$$f_T(t) = \begin{cases} f_X(h^{-1}(t)) \left| \frac{d}{dt} h^{-1}(t) \right|, \textit{if } t \in E_T, \\ 0, \textit{otherwise.} \end{cases}$$

Problem 5. *Let X be an RV with p.d.f.*

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

and let $T = X^2$. Show that T is an RV of continuous type and find its p.d.f. Also, find the p.d.f. of T by computing its c.d.f.

Solution. Clearly T is an RV and $E_X = E_T = (0, \infty)$. Also, $h(x) = x^2$, $x \in E_X$ is strictly increasing on E_X with inverse $h^{-1}(x) = \sqrt{x}$, $x \in E_T$. From the above theorem, T is a continuous type RV with p.d.f.

$$\begin{aligned} f_T(t) &= \begin{cases} f_X(\sqrt{t}) \left| \frac{d}{dt}(\sqrt{t}) \right|, \text{if } t > 0, \\ 0, \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{e^{-\sqrt{t}}}{2\sqrt{t}}, \text{if } t > 0, \\ 0, \text{otherwise.} \end{cases} \end{aligned}$$

Now, let us find the c.d.f. of T.

$$\begin{aligned} F_T(t) &= P(T \le t) \\ &= P(X^2 \le t) \\ &= \begin{cases} 0, \text{if } t \le 0, \\ P(-\sqrt{t} \le X \le \sqrt{t}), \text{if } t > 0 \end{cases} \\ &= \begin{cases} 0, \text{if } t \le 0, \\ \int_0^{\sqrt{t}} f_X(x) dx, \text{if } t > 0 \end{cases} \\ &= \begin{cases} 0, \text{if } t \le 0, \\ -e^{-\sqrt{t}}, \text{if } t > 0. \end{cases} \end{aligned}$$

We see that F_T is differentiable everywhere except $t = 0$. Therefore, the p.d.f. of T is given by

$$\begin{aligned} f_T(t) &= \begin{cases} (-e^{-\sqrt{t}})', \text{if } t > 0, \\ 0, \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{e^{-\sqrt{t}}}{2\sqrt{t}}, \text{if } t > 0, \\ 0, \text{otherwise.} \end{cases} \end{aligned}$$

Expectation, Variance and Standard Deviation

Measure of central tendency: gives an idea about the central value of the probability distribution around which values of the random variable are clustered. Mean, median and mode are three commonly used measures of central tendency.

Measure of dispersion: Measures of central tendency give us the idea about the location of only central part of the distribution. Other measures are often needed to describe a probability distribution. A probability distribution (or the corresponding random variable) is said to have a high dispersion if its support contains many values that are significantly higher or lower than the mean or median value. Some of the commonly used measures of dispersion are standard deviation, quartile deviation (or semi-inter-quartile range) and coefficient of variation.

Definition 1. (1) *Let X be a discrete random variable with p.m.f. f_X and support E_X. We say that the expected value of X or mean of X or expectation of X (denoted by $E(X)$) is finite and equals*

$$E(X) = \sum_{x \in E_X} x f_X(x)$$

provided the series $\sum_{x \in E_X} |x| f_X(x)$ is convergent, i.e., $\sum_{x \in E_X} |x| f_X(x) < \infty$.

(2) *Let X be a continuous random variable with p.d.f. f_X. We say that the expected value of X or mean of X or expectation of X (denoted by $E(X)$) is finite and equals*

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

provided the integral $\int_{-\infty}^{\infty} |x| f_X(x) dx$ is convergent, i.e., $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$.

Remark 2. (1) *Let X be a discrete random variable with finite support E_X and the p.m.f. f_X. Then $\sum_{x \in E_X} |x| f_X(x) < \infty$. Hence, $E(X)$ is finite.*

(2) *Let X be a continuous random variable with support $E_X \subseteq [-a, a]$ and p.d.f. f_X, for some $0 < a < \infty$. Then $\int_{-\infty}^{\infty} |x| f_X(x) dx = \int_{-a}^{a} |x| f_X(x) dx \leq a \int_{-a}^{a} f_X(x) dx = a$. Hence, $E(X)$ is finite.*

(3) *Let X be a continuous random variable with the d.f. F_X and p.d.f. f_X. Then*

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x)dx \\
&= \int_{-\infty}^{0} x f_X(x)dx + \int_{0}^{\infty} x f_X(x)dx \\
&= -\int_{-\infty}^{0}\int_{x}^{0} f_X(x)dtdx + \int_{0}^{\infty}\int_{0}^{x} f_X(x)dtdx \\
&= -\int_{-\infty}^{0}\int_{-\infty}^{t} f_X(x)dxdt + \int_{0}^{\infty}\int_{t}^{\infty} f_X(x)dxdt \text{ (by the change of order of integration)} \\
&= -\int_{-\infty}^{0} P(\{X < t\})dt + \int_{0}^{\infty} P(\{X > t\})dt \\
&= \int_{0}^{\infty}(1 - F_X(t))dt - \int_{-\infty}^{0} F_X(t-)dt
\end{aligned}
$$

Hence $E(X)$ does not depend on the version of p.d.f. although p.d.f. may not be unique.

Example 3. *Let X be a random variable with p.m.f.*

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

Show that the expected value of X is finite and find its value.

Solution: The support $E_X = \{1, 2, 3, \cdots\}$. By the ratio test, we can see that the infinite series $\sum_{x \in E_X} |x| f_X(x) = \sum_{n=1}^{\infty} \frac{n}{2^n}$ is convergent. Hence, the expected value of X is finite.

Now $E(X) = \sum_{x \in E_X} x f_X(x) = \sum_{n=1}^{\infty} \frac{n}{2^n} = \lim_{n\to\infty} s_n$, where the partial sum $s_n = 2[1 - \frac{1}{2^n} - \frac{n}{2^{n+1}}]$. This implies that $E(X) = 2$.

Example 4. *Let X be a random variable with p.d.f. $f_X(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty$ Show that the expected value of X is not finite.*

Solution: We have $\int_{-\infty}^{\infty} |x| f_X(x)dx = \int_{-\infty}^{\infty} |x| \frac{1}{\pi(1+x^2)}dx = \int_{0}^{\infty} \frac{2x}{\pi(1+x^2)}dx$. Since the improper integral $\int_{0}^{\infty} \frac{2x}{\pi(1+x^2)}dx$ is not convergent, the expected value of X is not finite.

Theorem 5. (1) *Let X be a random variable of discrete type with the support E_X and the p.m.f. f_X. Let $h : \mathbb{R} \longrightarrow \mathbb{R}$ be a function such that $h^{-1}(A) \in \mathbb{B}_{\mathbb{R}}$, for every $A \in \mathbb{B}_{\mathbb{R}}$. Then*

$$
E(h(X)) = \sum_{x \in E_X} h(x) f_X(x);
$$

(2) *Let X be a random variable of continuous type with p.d.f. f_X. Let $h : \mathbb{R} \longrightarrow \mathbb{R}$ be a function such that $h^{-1}(A) \in \mathbb{B}_{\mathbb{R}}$, for every $A \in \mathbb{B}_{\mathbb{R}}$. Then*

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x);$$

(3) *If, for real constants a and b with $a \leq b$, $P(\{a \leq X \leq b\}) = 1$, then $a \leq E(X) \leq b$;*

(4) *Let $h_i : \mathbb{R} \longrightarrow \mathbb{R}$ be a function such that $h_i^{-1}(A) \in \mathbb{B}_{\mathbb{R}}$, for every $A \in \mathbb{B}_{\mathbb{R}}$, for $i = 1, 2$. If $P(\{h_1(X) \leq h_2(X)\}) = 1$, then $E(h_1(X)) \leq E(h_2(X))$, provided the involved expectations are finite;*

(5) *Let $h_i : \mathbb{R} \longrightarrow \mathbb{R}$ be a function such that $h_i^{-1}(A) \in \mathbb{B}_{\mathbb{R}}$, for every $A \in \mathbb{B}_{\mathbb{R}}$, for $i = 1, 2, \cdots, m$. Then*

$$E(\sum_{i=1}^{m} h_i(X)) = \sum_{i=1}^{m} E(h_i(X)),$$

provided the involved expectations are finite;

(6) *If $P(\{X \geq 0\}) = 1$ and $E(X) = 0$, then $P(\{X = 0\}) = 1$.*

Definition 6. *Let X be a random variable.*

(1) *For $r \in \{1, 2, \cdots\}$, $\mu'_r = E(X^r)$, provided it is finite, is called the r-th moment of the random variable X;*

(2) *For $r \in \{1, 2, \cdots\}$, $E(|X|^r)$, provided it is finite, is called the r-th absolute moment of the random variable X;*

(3) *For $r \in \{1, 2, \cdots\}$, $\mu_r = E((X - \mu'_1)^r)$, provided it is finite, is called the r-th central moment of the random variable X;*

(4) *$\mu_2 = E((X - \mu'_1)^2) = E((X - E(X))^2)$, provided it is finite, is called the variance of the random variable X. The variance of the random variable X is denoted by $Var(X)$. The quantity $\sigma = \sqrt{Var(X)} = \sqrt{E((X - E(X))^2)}$ is called the standard deviation of the random variable X.*

Proposition 7. *Let X be a random variable with finite first two moments and $\mu = E(X)$. Then*

(1) *For real constants a and b, $E(aX + b) = aE(X) + b$;*

(2) *$Var(X) = E(X^2) - (E(X))^2$;*

(3) *$Var(X) \geq 0$. Moreover, $Var(X) = 0$ if and only if, $P(\{X = \mu\}) = 1$;*

(4) *$E(X^2) \geq (E(X))^2$ (Cauchy- Schwarz inequality);*

(5) *For real constants a and b, $Var(aX + b) = a^2 Var(X)$.*

Proof. (1) Consider the function $h : \mathbb{R} \longrightarrow \mathbb{R}$, defined by $h(x) = ax + b$, for all $x \in \mathbb{R}$. Since h is a continuous function, $h(X) = aX + b$ is random variable.

Suppose X is a discrete type random variable with the p.m.f. f_X and the support E_X. Then by Theorem 5(1),

$$\begin{aligned} E(h(X)) &= E(aX+b) \\ &= \sum_{x\in E_X} h(x) f_X(x) \\ &= \sum_{x\in E_X} (ax+b) f_X(x) \\ &= a\sum_{x\in E_X} x f_X(x) + b\sum_{x\in E_X} f_X(x) \\ &= aE(X) + b \text{ (since } \sum_{x\in E_X} f_X(x) = 1) \end{aligned}$$

Similarly, we can prove it for continuous case.

(2)

$$\begin{aligned} Var(X) &= E((X-\mu)^2) \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \text{ (by using (1) and Theorem 5(4))} \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

(3) Since $P(\{(X-\mu)^2 \geq 0\}) = P(\mathcal{S}) = 1$, using Theorem 5(3), $E(0) \leq E((X-\mu)^2)$. So, $Var(X) \geq 0$.

Also, using Theorem 5(5), if $Var(X) = E((X-\mu)^2) = 0$, then $P(\{(X-\mu)^2 = 0\}) = 1$, i.e., $P(\{X=\mu\}) = 1$. Conversely, if $P(\{X=\mu\}) = 1$, then $E(X) = \mu$ and $E(X^2) = \mu^2$ (as support of X is $\{\mu\}$). Now, $Var(X) = E(X^2) - (E(X))^2 = \mu^2 - \mu^2 = 0$.

(4) Since $Var(X) \geq 0$ and $Var(X) = E(X^2) - (E(X))^2$, $E(X^2) \geq (E(X))^2$.

(5) Let $Y = aX + b$. Then, (1), we have $E(Y) = aE(X) + b$. So, $Y - E(Y) = a(X - E(X))$.

$$\begin{aligned} Var(aX+b) &= Var(Y) \\ &= E((Y-E(Y))^2) \\ &= E(a^2(X-E(X))^2) \\ &= a^2 E((X-E(X))^2) \\ &= a^2 Var(X) \end{aligned}$$

□

Example 8. *Let X be a random variable with p.m.f.*

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

Find the expectation and variance of X.

Solution: $E(X) = \sum_{x\in E_X} x f_X(x) = \frac{7}{2}$.

$E(X^2) = \sum_{x \in E_X} x^2 f_X(x) = \frac{91}{6}$.

Hence, $Var(X) = E(X^2) - (E(X))^2 = \frac{25}{12}$.

Example 9. *Let X be a random variable with p.m.f.*

$$f_X(x) = \begin{cases} 0.2, & \text{if } x = 0 \\ 0.5, & \text{if } x = 1 \\ 0.3, & \text{if } x = 2 \\ 0, & \text{otherwise} \end{cases}$$

(1) *Find the p.m.f. of $Y = X^2$ and hence find expectation of Y.*
(2) *Find expectation of Y directly.*

Solution:

(1) The support of Y is $E_Y = \{x^2 \mid x \in E_X\} = \{0, 1, 4\}$. Now, the p.m.f. of $Y = X^2$ is

$$f_Y(y) = \begin{cases} 0.2, & \text{if } y = 0 \\ 0.5, & \text{if } y = 1 \\ 0.3, & \text{if } y = 4 \\ 0, & \text{otherwise} \end{cases}$$

Hence, $E(Y) = \sum_{y \in E_Y} y f_Y(y) = 1.7$

(2) $E(X^2) = \sum_{x \in E_X} x^2 f_X(x) = 1.7$

Example 10. *Let X be a random variable with p.d.f.*

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

(1) *Find the p.d.f. of $Y = X^3$ and hence find expectation of Y.*
(2) *Find expectation of Y directly.*

Solution:

(1) The c.d.f. of Y is

$$F_Y(y) = P(Y \leq y) = \begin{cases} 0, & \text{if } y \leq 0 \\ \int_0^{y^{1/3}} dx, & \text{if } 0 < y < 1 \\ \int_0^1 dx, & \text{if } y \geq 1 \end{cases}$$

Thus

$$F_Y(y) = \begin{cases} 0, & \text{if } y \leq 0 \\ y^{1/3}, & \text{if } 0 < y < 1 \\ 1, & \text{if } y \geq 1 \end{cases}$$

Hence the p.d.f. of Y is

$$f_Y(y) = \begin{cases} \frac{1}{3} y^{-2/3}, & \text{if } 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Now, $E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_0^1 y f_Y(y)dy = \int_0^1 \frac{1}{3}y^{1/3}dy = \frac{1}{4}$.

(2) $E(Y) = E(X^3) = \int_{-\infty}^{\infty} x^3 f_X(x)dx = \int_0^1 x^3 dx = \frac{1}{4}$.

Example 11. *Let X be a random variable with p.d.f.*

$$f_X(x) = \begin{cases} \frac{1}{2}, & \text{if } -2 < x < -1 \\ \frac{x}{9}, & \text{if } 0 < x < 3 \\ 0, & \text{otherwise} \end{cases}$$

(1) *If $Y_1 = \max(X, 0)$, find the mean and variance of Y_1.*
(2) *If $Y_2 = 2X + 3e^{-\max(X,0)} + 4$, find the mean Y_2.*

Solution:

(1) Let $h : \mathbb{R} \longrightarrow \mathbb{R}$ be a function defined by $h(x) = \max(x, 0)$, for all $x \in \mathbb{R}$. Then $Y_1 = h(X)$. For $r > 0$, $E(Y_1^r) = \int_{-\infty}^{\infty} (\max(x, 0))^r f_X(x)dx = \int_0^3 \frac{x^{r+1}}{9}dx = \frac{3^r}{r+2}$. Hence $E(Y_1) = 1$ and $Var(Y_1) = E(Y_1^2) - (E(Y_1))^2 = \frac{9}{4} - 1 = \frac{5}{4}$.

(2) We have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)dx \\ &= \int_{-2}^{-1} \frac{x}{2}dx + \int_0^3 \frac{x^2}{9}dx \\ &= \frac{1}{4} \end{aligned}$$

and

$$\begin{aligned} E(e^{-\max(X,0)}) &= \int_{-\infty}^{\infty} e^{-\max(x,0)} f_X(x)dx \\ &= \int_{-2}^{-1} \frac{1}{2}dx + \int_0^3 \frac{xe^{-x}}{9}dx \\ &= \frac{11 - 8e^{-3}}{18}. \end{aligned}$$

Therefore, $E(Y_2) = 2E(X) + 3E(e^{-\max(X,0)}) + 4 = \frac{19-4e^{-3}}{3}$

Definition 12. *Let X be a random variable with the d.f. F_X. A real number x satisfying*

$$F_X(x-) \leq \frac{1}{2} \leq F_X(x), \text{ i.e., } P(\{X < x\}) \leq \frac{1}{2} \leq P(\{X \leq x\})$$

is called a median of X. A median is not necessarily unique.

Moment generating function and Moment Inequalities

Let X be a random variable and let $A = \{t \in \mathbb{R} \mid E(e^{tX}) \text{ is finite}\}$. The function $M_X : A \longrightarrow \mathbb{R}$, defined by

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

is known as the moment generating function (m.g.f.) of the random variable X if $E(e^{tX})$ is finite on an interval $(-a, a) \subseteq A$, for some $a > 0$.

Theorem 1. *Let X be a random variable with the moment generating function (m.g.f.) M_X that is finite on an interval $(-a, a)$, for some $a > 0$. Then*

(1) *for each $r \in \{1, 2, \cdots\}$, $M_X^{(r)}(t)$ exists on $(-a, a)$, and for each $r \in \{1, 2, \cdots\}$, $\mu'_r = E(X^r)$ is finite and is equal to $\mu'_r = E(X^r) = M_X^{(r)}(0)$, where $M_X^{(r)}(t) = \frac{d^r M_X(t)}{dt^r}$;*

(2) *$M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu'_r$, $t \in (-a, a)$.*

Example 2. *Let X be a random variable with the p.m.f.*

$$f_X(k) = \begin{cases} \frac{6}{\pi^2 k^2}, & \text{if } k \in \{1, 2, \cdots\} \\ 0, & \text{otherwise.} \end{cases}$$

Then $\frac{1}{\pi^2} \sum_{k=1}^{\infty} \frac{e^{tk}}{k^2}$ is not convergent for every $t > 0$. Thus the moment generating function (m.g.f.) of the random variable X does not exist.

Example 3. *Let X be a random variable with p.d.f.*

$$f_X(x) = \begin{cases} \frac{1}{2}e^{-x/2}, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Now, the moment generating function (m.g.f.) of the random variable X

$$M_X(t) = E(e^{tX}) = \frac{1}{2}\int_0^{\infty} e^{(t-1/2)x}dx = \frac{1}{1-2t}, \ t < \frac{1}{2}.$$

Also $M_X^{(1)}(t) = \frac{2}{(1-2t)^2}$ and $M_X^{(2)}(t) = \frac{8}{(1-2t)^3}$, $t < \frac{1}{2}$. It follows that

$$E(X) = 2, E(X^2) = 8, \text{ and } Var(X) = 4$$

Proposition 4. *Let X be a continuous random variable that takes only non-negative values with a p.d.f. f_X. Then there exists a p.d.f. g_X of X such that $g_X(x) = 0$, for $x < 0$.*

Proof. Since X takes only non-negative values, $P(\{X < 0\}) = 0$. Then the function $g_X : \mathbb{R} \longrightarrow \mathbb{R}$, defined by,

$$g_X(x) = \begin{cases} f_X(x), & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

is a p.d.f. of X (verify!). □

Theorem 5. ***(Markov's Inequality)*** *If X is random variable that takes only non-negative values, then for any $a > 0$,*

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

Proof. Suppose X is a continuous random variable with a p.d.f. f of X such that $f(x) = 0$, for $x < 0$. Then

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_{0}^{\infty} xf(x)dx \\ &= \int_{0}^{a} xf(x)dx + \int_{a}^{\infty} xf(x)dx \\ \Rightarrow E(X) &\geq \int_{a}^{\infty} xf(x)dx \\ \Rightarrow E(X) &\geq a\int_{a}^{\infty} f(x)dx \\ \Rightarrow E(X) &\geq aP(X \geq a) \\ \Rightarrow P(\{X \geq a\}) &\leq \frac{E(X)}{a}. \end{aligned}$$

□

General form of Markov Inequality: Suppose that $E(|X|^r) < \infty$, for some $r > 0$. Then, for any any $a > 0$,

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

Corollary 6. ***(Chebyshev Inequality)*** *Suppose that random variable has finite first two moments. If $\mu = E(X)$ and $\sigma^2 = Var(X)$. Then, for any any $a > 0$,*

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

Proof. We have $P(\{|X-\mu| \geq a\}) = P(\{|X-\mu|^2 \geq a^2\})$ (verify it). Using the Markov Inequality on the random variable $|X - \mu|^2$, we have

$$\begin{aligned} P(\{|X - \mu|^2 \geq a^2\}) &\leq \frac{E((X-\mu)^2)}{a^r} = \frac{\sigma^2}{a^r} \\ \Rightarrow P(\{|X - \mu| \geq a\}) &\leq \frac{\sigma^2}{a^r}. \end{aligned}$$

□

Example 7. *Let X be a random variable with the p.m.f.*

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

Then $E(X) = \sum\limits_{x \in S_X} xf_X(x) = 0$ and $E(X^2) = \sum\limits_{x \in S_X} x^2 f_X(x) = \frac{1}{4}$. Therefore, using the Markov Inequality, we have

$$P(\{|X| \geq 1\}) \leq \frac{E(X^2)}{1} = \frac{1}{4}.$$

The exact probability is

$$P(\{|X| \geq 1\}) = P(\{X \in \{-1, 1\}\}) = \frac{1}{4}.$$

Definition 8. *A random variable X is said to have a symmetric distribution about a point $\mu \in \mathbb{R}$ if $P(\{X \leq \mu + x\}) = P(\{X \geq \mu - x\})$, $\forall\ x \in \mathbb{R}$, i.e., $F_X(\mu - x) + F_X(\mu + x) = 1$, $\forall\ x \in \mathbb{R}$.*

Remark 9. *Let X be a random variable having p.d.f./p.m.f. f_X and $\mu \in \mathbb{R}$. Then the distribution of X is symmetric about μ if and only if $f_X(\mu - x) = f_X(\mu + x)$, $\forall\ x \in \mathbb{R}$.*

Bernoulli, Binomial and Uniform Distributions

Let $(\mathcal{S}, \Sigma, P)$ be a probability space corresponding to a random experiment $\mathcal{E}$.

- Each repetition of the random experiment $\mathcal{E}$ will be called a trial.
- We say that a collection of trials forms a collection of independent trials if any collection of corresponding events forms a collection of independent events.

1. Bernoulli Distribution

A random experiment is said to be a Bernoulli experiment if its each trial results in just two possible outcomes, labeled as success (s) and failure (f). Each repetition of a Bernoulli experiment is called a Bernoulli trial. For example, consider a sequence of random rolls of a fair dice. In each roll of the dice a person bets on occurrence of upper face with six dots. Let the event of occurrence of upper face with six dots be denoted by E. Here, in each trial, one is only interested in the occurrence or non-occurrence of the event E. In such situations, the occurrence of event E will be label as a success and the non-occurrence of event E will be label as a failure.

For a Bernoulli trial, the sample space is $\mathcal{S} = \{s, f\}$, the event space is $\Sigma = \mathcal{P}(\mathcal{S})$ and the probability function is $P : \Sigma \longrightarrow \mathbb{R}$ defined by $P(\{s\}) = p$, $P(\{f\}) = 1 - p$, $P(\{\emptyset\}) = 0$ and $P(\{\mathcal{S}\}) = 1$, where $p \in (0, 1)$ is a fixed real number and it is the probability of success of the trial. Define the random variable $X : \mathcal{S} \longrightarrow \mathbb{R}$ by

$$X(w) = \begin{cases} 1, \text{ if } w = s \\ 0, \text{ if } w = f \end{cases}$$

Then the r.v. X is of discrete type with the support $E_X = \{0, 1\}$ and the p.m.f.

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

The random variable X is called a Bernoulli random variable and the distribution with p.m.f. (1) is called a Bernoulli distribution with success probability $p \in (0, 1)$.

The d.f. of X is given by

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

Now, the expectation of X is $E(X) = \sum\limits_{x \in \{0,1\}} x f_X(x) = p$ and $E(X^2) = \sum\limits_{x \in \{0,1\}} x^2 f_X(x) = p$. Thus the variance is $Var(X) = p - p^2 = p(1-p)$. Also the moment generating functions is

$$M_X(t) = E(e^{tX}) = \sum_{x \in \{0,1\}} e^{tx} f_X(x) = p(e^t - 1) + 1, \ \forall\ t \in \mathbb{R}$$

.

2. Binomial Distribution

Consider a sequence of n independent Bernoulli trials with probability of success (s) in each trial being $p \in (0,1)$. In this case, the sample space is $\mathcal{S} = \{(w_1, w_2, \ldots, w_n) \mid w_i \in \{s, f\}, i = 1, 2, \ldots, n\}$, where w_i represents the outcome of the i-th Bernoulli trial and the event space is $\Sigma = \mathcal{P}(\mathcal{S})$. Define the random variable $X : \mathcal{S} \longrightarrow \mathbb{R}$ by

$$X((w_1, w_2, \ldots, w_n)) = \text{number of successes among } w_1, w_2, \ldots, w_n$$

Clearly, Im X $= \{0, 1, 2, \cdots, n\}$ and $P(\{X = x\}) = 0$, if $x \notin \{0, 1, 2, \cdots, n\}$. For $x \in \{0, 1, 2, \cdots, n\}$

$$\begin{aligned} P(\{X = x\}) &= P(\{(w_1, w_2, \ldots, w_n) \in S \mid X(w_1, w_2, \ldots, w_n) = x\}) \\ &= \sum_{(w_1, w_2, \ldots, w_n) \in S_x} P((w_1, w_2, \ldots, w_n)), \end{aligned}$$

where $S_x = \{(w_1, w_2, \ldots, w_n) \mid x \text{ of } w_i's \text{ are } s \text{ and remaining } n - x \text{ of } w_i's \text{ are } f\}$.

For $x \in \{0, 1, 2, \cdots, n\}$ and $(w_1, w_2, \ldots, w_n) \in S_x$,

$$P((w_1, w_2, \ldots, w_n)) = p^x (1-p)^{n-x},$$

since trials are independent and $P(\{s\}) = p$ & $P(\{f\}) = 1 - p$. Therefore, $x \in \{0, 1, 2, \cdots, n\}$,

$$P(\{X = x\}) = \sum_{(w_1, w_2, \ldots, w_n) \in S_x} p^x (1-p)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x}.$$

Thus the r.v. X is of discrete type with support $E_X = \{0, 1, 2, \cdots, n\}$ and p.m.f.

$$f_X(x) = P(\{X = x\}) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{if } x \in \{0, 1, 2, \cdots, n\} \\ 0, & \text{otherwise} \end{cases} \quad . \tag{2}$$

The random variable X is called a Binomial random variable with n trials and success probability $p \in (0,1)$ and it is written as $X \sim \text{Bin}(n, p)$. The probability distribution with the p.m.f. (2) is called a Binomial distribution with n trials and success probability $p \in (0,1)$. It is clear that $\sum\limits_{x \in E_X} f_X(x) = \sum\limits_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1$

Now, the expectation of $X \sim \text{Bin}(n,p)$ is

$$\begin{aligned}
E(X) &= \sum_{x \in E_X} x f_X(x) \\
&= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} \frac{xn!}{(n-x)x!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(n-x)(x-1)!} p^x (1-p)^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(n-x)(x-1)!} p^{(x-1)} (1-p)^{n-x} \\
&= np \sum_{x=0}^{n-1} \binom{n-1}{x} p^x (1-p)^{n-1-x} \\
&= np(p+(1-p))^{(n-1)} = np
\end{aligned}$$

Now, the moment generating function of $X \sim \text{Bin}(n,p)$ is

$$\begin{aligned}
M_X(t) &= E(e^{tX}) \\
&= \sum_{x \in E_X} e^{tx} f_X(x) \\
&= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\
&= (pe^t + (1-p))^n, \ t \in \mathbb{R}
\end{aligned}$$

Therefore,

$M_X^{(1)}(t) = npe^t(pe^t + (1-p))^{(n-1)}, \ t \in \mathbb{R}$;
$M_X^{(2)}(t) = npe^t(pe^t + (1-p))^{(n-1)} + n(n-1)p^2e^{2t}(pe^t + (1-p))^{(n-2)}, \ t \in \mathbb{R}$;
$E(X) = M_X^{(1)}(0) = np$;
$E(X^2) = M_X^{(2)}(0) = np + n(n-1)p^2$;
and $Var(X) = E(X^2) - (E(X))^2 = np(1-p)$.

Example 1. *Four fair coins are flipped. If the outcomes are assumed independent, what is the probability that two heads and two tails are obtained?*

Solution: Let us label the occurrence of a head in a trial as success and label the occurrence of a tail in a trial as failure. Let X be the number of successes (i.e. heads) that appear. Then $X \sim \text{Bin}(4, \frac{1}{2})$. Hence the required probability is $P(X=2) = \binom{4}{2}(\frac{1}{2})^2(\frac{1}{2})^2 = \frac{3}{8}$.

Example 2. *A fair dice is rolled six times independently. Find the probability that on two occasions we get an upper face with* 2 *or* 3 *dots.*

Solution: Let us label the occurrence of an upper face having 2 or 3 dots as success and label the occurrence of any other face as failure. Let X be the number of occasions on which we get success (i.e., an upper face having 2 or 3 dots). Then $X \sim \text{Bin}(6, \frac{1}{3})$. Hence the required probability is $P(X = 2) = \binom{6}{2}(\frac{1}{3})^2(\frac{2}{3})^4 = \frac{80}{243}$.

3. Discrete Uniform Distribution

For a given positive integer $N(\geq 2)$ and real numbers $x_1 < x_2 < \cdots < x_N$, a random variable X of discrete type is said to follow a discrete uniform distribution on the set $\{x_1, x_2, \ldots, x_N\}$ (written as $X \sim U(\{x_1, x_2, \ldots, x_N\})$) if the support of X is $E_X = \{x_1, x_2, \ldots, x_N\}$ and its p.m.f. is given by

$$f_X(x) = P(\{X = x\}) = \begin{cases} \frac{1}{N}, \text{ if } x \in E_X = \{x_1, x_2, \ldots, x_N\} \\ 0, \text{ otherwise} \end{cases}$$

.

Now, for $r \in \{1, 2, \cdots\}$, $E(X^r) = \frac{1}{N}\sum_{i=1}^{N} x_i^r$. Therefore, the mean $E(X) = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $Var(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(X))^2$. Also the m.g.f. is $M_X(t) = E(e^{tX}) = \frac{1}{N}\sum_{i=1}^{N} e^{tx_i}$, $t \in \mathbb{R}$.

Now, suppose that $X \sim U(\{1, 2, \ldots, N\})$. Then

$$E(X) = \frac{1}{N}\sum_{i=1}^{N} i = \frac{N+1}{2},$$

$$E(X^2) = \frac{1}{N}\sum_{i=1}^{N} i^2 = \frac{(N+1)(2N+1)}{6},$$

$$Var(X) = E(X^2) - (E(X))^2 = \frac{N^2-1}{12}.$$

Also the m.g.f. of $X \sim U(\{1, 2, \ldots, N\})$ is

$$M_X(t) = E(e^{tX}) = \frac{1}{N}\sum_{i=1}^{N} e^{it} = \begin{cases} \frac{e^t(e^{Nt}-1)}{e^t-1}, \text{ if } t \neq 0 \\ 1, \text{ if } t = 0 \end{cases}$$

Negative Binomial and Geometric Distribution

1. Negative Binomial Distribution

Let $r \in \mathbb{N}$. Suppose that we keep performing independent Bernoulli trials until the r-th success is observed. Further suppose that the probability of success in each trial is $p \in (0,1)$. Thus, the sample space is

$\mathcal{S} = \{(w_1, w_2, \ldots, w_n) : n \in \{r, r+1, \ldots\}, w_n = s, w_i \in \{s, f\}, i = 1, 2, \ldots, n-1; r-1 \text{ of } w_1, w_2, \ldots, w_{n-1} \text{ are } s \text{ and remaining } n-r \text{ of } w_1, w_2, \ldots, w_{n-1} \text{ are } f\}$.

Note: To find r-th success, we have to perform Bernoulli trials at least r-times. Thus, $(w_1, w_2, \ldots, w_n) \in S$ corresponds to one of $\binom{n-1}{r-1}$ ways in which the r-th success is obtained in the n-th Bernoulli trials $w_n = s$ and the first $n-1$ Bernoulli trials result in $r-1$ successes and $n-r$ failures.

Define the r.v. $X : \mathcal{S} \longrightarrow \mathbb{R}$ by

$$\begin{aligned} X((w_1, w_2, \ldots, w_n)) &= n - r \\ &= \text{number of failures preceding the } r\text{-th success} \end{aligned}$$

Clearly, for $x \notin \{0, 1, 2, \cdots\}$, $P(\{X = x\}) = 0$. Also, for $x \in \{0, 1, 2, \cdots\}$, event $\{X = x\}$ occurs if and only if the $(r+x)$-th trial results in success and the first $(r+x-1)$ Bernoulli trials result in $r-1$ successes and x failures are observed. Since the trials are independent, for $x \in \{0, 1, 2, \cdots\}$, we have

$$P(\{X = x\}) = p_1 p_2,$$

where p_1 is the probability of observing $(r-1)$ successes in the first $(r+x-1)$ independent Bernoulli trials and p_2 is the probability of getting the success on the $(r+x)$-th trial. Clearly, $p_2 = p$, and using the property of Binomial distribution

$$p_1 = \binom{r+x-1}{r-1} p^{r-1}(1-p)^x.$$

Therefore, $x \in \{0, 1, 2, \cdots\}$,

$$P(\{X = x\}) = \binom{r+x-1}{r-1} p^r (1-p)^x.$$

Thus, the r.v. X is of discrete type with support $E_X = \{0, 1, 2, \cdots\}$ and p.m.f.

$$f_X(x) = P(\{X = x\}) = \begin{cases} \binom{r+x-1}{r-1} p^r (1-p)^x, & \text{if } x \in \{0, 1, 2, \cdots\} \\ 0, & \text{otherwise} \end{cases} . \tag{1}$$

The random variable X is called a Negative Binomial random variable with r successes and success probability $p \in (0,1)$ and it is written as $X \sim \text{NB}(r, p)$. The probability distribution with the p.m.f. (1) is called a Negative Binomial distribution with r successes and success probability $p \in (0,1)$.

Remark 1. (1) *It is easy to see that the series* $\sum_{x=0}^{\infty} \binom{r+x-1}{r-1} t^x$ *is an absolutely convergent series, for* $|t| < 1$ *and* $\sum_{x=0}^{\infty} \binom{r+x-1}{r-1} t^x = (1-t)^{-r}$, *for* $|t| < 1$. *Thus,* $\sum_{x \in E_X} f_X(x) = p^r \sum_{x=0}^{\infty} \binom{r+x-1}{r-1} (1-p)^x = p^r (1-(1-p))^{-r} = 1$.

(2) *Consider a sequence of independent Bernoulli trials with probability of success in each trial being* p. *Let* Z *denote the number of trials required to get the* r*-th success, where* $r \in \mathbb{N}$ *and* $X = Z - r$. *Then* $X \sim NB(r, p)$.

Now, the m.g.f. of $X \sim \text{NB}(r, p)$ is

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \sum_{x \in E_X} e^{tx} f_X(x) \\ &= \sum_{x=0}^{\infty} e^{tx} \binom{r+x-1}{r-1} p^r q^x, \text{ where } q = 1-p \\ &= \sum_{x=0}^{\infty} \binom{r+x-1}{r-1} p^r (qe^t)^x \\ &= p^r (1-qe^t)^{-r}, \ |qe^t| < 1 \\ &= \left(\frac{p}{1-qe^t}\right)^r, \ |t| < -\ln q \end{aligned}$$

Therefore,

$$\begin{aligned} &M_X^{(1)}(t) = p^r \{rqe^t(1-qe^t)^{-r-1}\}, \ |t| < -\ln q; \\ &M_X^{(2)}(t) = p^r \{rqe^t(1-qe^t)^{-r-1} + r(r+1)(qe^t)^2(1-qe^t)^{-r-2}\}, \ |t| < -\ln q; \\ &E(X) = M_X^{(1)}(0) = \frac{rq}{p}; \\ &E(X^2) = M_X^{(2)}(0) = \frac{r(r+1)q^2}{p^2} + \frac{rq}{p}; \\ &\text{and } Var(X) = E(X^2) - (E(X))^2 = \frac{rq}{p^2}, \text{ where } q = 1-p. \end{aligned}$$

Example 2. *A person repeatedly rolls a fair dice independently until an upper face with two or three dots is observed twice. Find the probability that the person would require eights rolls to achieve this.*

Solution: In each trial, let us label the outcome of observing an upper face with two or three dots as success and observing any other outcome as a failure. Hence success probability in each trial is $\frac{1}{3}$. Let Z denote the number of trials required to get the second success and $X = Z - 2$. Then $X \sim \text{NB}(2, \frac{1}{3})$. Therefore, the required probability is

$$P(\{Z = 8\}) = P(\{X = 6\}) = \binom{7}{1}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^6 = \frac{448}{6561}.$$

Example 3. *A mathematician carries one matchbox each in his right and left pockets. When he wants a match, he selects the left pocket with probability p and the right pocket with probability $1-p$. Suppose that initially each box contains N matches. Consider the moment when the mathematician for the first time discovers that one of the match boxes is empty. Find the probability that at that moment the other box contains exactly k matches, where $k \in \{0, 1, 2, \cdots, N\}$.*

Solution: Let us identify success with the choice of the left pocket. The left pocket box will be empty at the moment when the right pocket box contains exactly k matches if and only if $N-k$ failures precede the $(N+1)$-th success. A similar arguments applies to the right pocket.

Now the required probability is

$$\begin{aligned}
p &= P(\text{the left pocket is found empty, the right pocket contains k matches}) \\
&\quad + P(\text{the right pocket is found empty, the left pocket contains k matches}) \\
&= \binom{N+1+N-k-1}{N+1-1} p^{N+1}(1-p)^{N-k} + \binom{N+1+N-k-1}{N+1-1}(1-p)^{N+1}p^{N-k} \\
&= \binom{2N-k}{N} p^{N+1}(1-p)^{N-k} + \binom{2N-k}{N}(1-p)^{N+1}p^{N-k}.
\end{aligned}$$

2. Geometric Distribution

An NB$(1,p)$ distribution is called a geometric distribution with success probability p and is denoted by Ge(p). In this case, the sample space is $\mathcal{S} = \{(w_1, w_2, \ldots, w_n) \mid w_n = s, w_1, w_2, \ldots, w_{n-1} \text{ are } f\}$ and the r.v. $X : \mathcal{S} \longrightarrow \mathbb{R}$ is defined as

$$X((w_1, w_2, \ldots, w_n)) = n - 1 = \text{number of failures proceeding the first success.}$$

Hence, the p.m.f. and d.f. of X are

$$f_X(x) = P(\{X = x\}) = \begin{cases} pq^x, \text{ if } x \in \{0, 1, 2, \cdots\} \text{ where } q = 1-p \\ 0, \text{ otherwise} \end{cases},.$$

and

$$\begin{aligned}
F_X(x) &= P(\{X \le x\}) \\
&= \begin{cases} 0, \text{ if } x < 0 \\ p\sum_{x=0}^{k} q^x, \text{ if } k \le x < k+1, \text{ where } k = 0, 1, \cdots \end{cases} \\
&= \begin{cases} 0, \text{ if } x < 0 \\ 1 - q^{k+1}, \text{ if } k \le x < k+1, \text{ where } k = 0, 1, \cdots \end{cases}.
\end{aligned}$$

respectively. Also, $M_X(t) = \frac{p}{1-qe^t}$, $|t| < -\ln q$, $E(X) = \frac{q}{p}$ and $Var(X) = \frac{q}{p^2}$.

Remark 4. *Suppose the r.v. $X : \mathcal{S} \longrightarrow \mathbb{R}$ is given by*

$$X((w_1, w_2, \ldots, w_n)) = n = \textit{number of trials to get the first success.}$$

Hence, the p.m.f. of X is

$$f_X(x) = P(\{X = x\}) = \begin{cases} pq^{x-1}, \textit{ if } x \in \{1, 2, \cdots\} \textit{ where } q = 1-p \\ 0, \textit{ otherwise} \end{cases},.$$

So, in this case the m.g.f. of X is

$$\begin{aligned}
M_X(t) &= E(e^{tX}) \\
&= \sum_{x \in E_X} e^{tx} f_X(x) \\
&= \sum_{x=1}^{\infty} e^{tx} p q^{x-1}, \\
&= pe^t \sum_{x=0}^{\infty} (qe^t)^x \\
&= \frac{pe^t}{1-qe^t}, \ |t| < -\ln q
\end{aligned}$$

Therefore,

$$M_X^{(1)}(t) = \frac{pe^t}{(1-qe^t)^2},\ |t| < -\ln q;$$

$$M_X^{(2)}(t) = \frac{(1-qe^t)^2 pe^t - 2pe^t(1-qe^t)(-qe^t)}{(1-qe^t)^4},\ |t| < -\ln q;$$

$$E(X) = M_X^{(1)}(0) = \frac{1}{p};$$

$$E(X^2) = M_X^{(2)}(0) = \frac{1+q}{p^2};$$

and $Var(X) = E(X^2) - (E(X))^2 = \frac{q}{p^2},$ *where* $q = 1-p.$

Hypergeometric and Poisson Distribution

1. Hypergeometric Distribution

Consider a population comprising of $N(\geq 2)$ units out of which $a(\in \{1, 2, \ldots, N-1\})$ are labeled as s (success) and $N-a$ are labeled as f (failure). A sample of size n is drawn from this population drawing one unit at a time. Let

$$X = \text{ number of successes in drawn sample}$$

Case I: Suppose draws are independent and sampling is with replacement (i.e., after each draw the drawn unit is replaced back into the population). Then we have a sequence of n independent Bernoulli trials with probability of success in each trial is $p = \frac{a}{N}$ and, therefore $X \sim \text{Bin}(n, \frac{a}{N})$.

Case II: Suppose sampling is without replacement (i.e., after each draw the drawn unit is not replaced back into the population).

$$P(\text{obtaining } s \text{ in first draw}) = \frac{a}{N};$$
$$P(\{\text{obtaining } s \text{ in second draw}\}) = \frac{a}{N}.\frac{a-1}{N-1} + \frac{N-a}{N}.\frac{a}{N-1} = \frac{a}{N};$$
$$P(\{\text{obtaining } s \text{ in third draw}\}) = \frac{a}{N}.\frac{a-1}{N-1}.\frac{a-2}{N-2} + \frac{a}{N}.\frac{N-a}{N-1}.\frac{a-1}{N-2}$$
$$+ \frac{N-a}{N}.\frac{a}{N-1}.\frac{a-1}{N-2} + \frac{N-a}{N}.\frac{N-a-1}{N-1}.\frac{a}{N-2} = \frac{a}{N};$$

In general, $P(\{\text{obtaining } s \text{ in } k-\text{th draw}\}) = \frac{a}{N}$.

Remark 1. *$P(\text{obtaining } s \text{ in first draw}) = \frac{a}{N}.\frac{a-1}{N-1}$ and*
$P(\text{obtaining } s \text{ in first draw})P(\text{obtaining } s \text{ in second draw}) = \frac{a}{N}.\frac{a}{N}$

This implies that the draws are not independent. Therefore, we cannot conclude that $X \sim Bin(n, \frac{a}{N})$.

For $P(\{X = x\}) \neq 0$, we have $0 \leq x \leq n, 0 \leq x \leq a$ and $0 \leq n - x \leq N - a$. Thus $P(\{X = x\}) \neq 0$, $x \in \{\max(0, n-N+a), \ldots, \min(n, a)\}$. Therefore the r.v. X is of discrete type with support $E_X = \{\max(0, n-N+a), \ldots, \min(n, a)\}$ and p.m.f.

$$(1) \quad f_X(x) = P(\{X = x\}) = \begin{cases} \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}, & \text{if } x \in \{\max(0, n-N+a), \ldots, \min(n, a)\} \\ 0, & \text{otherwise} \end{cases}$$

The random variable X is called a Hypergeometric random variable and it is written as $X \sim \text{Hyp}(a, n, N)$. The probability distribution with the p.m.f. (1) is called a Hypergeometric distribution. Also, we have

$$(2) \quad \sum_{x=\max(0,n-N+a)}^{\min(n,a)} \binom{a}{x}\binom{N-a}{n-x} = \binom{N}{n}$$

Now, the expectation of $X \sim \text{Hyp}(a, n, N)$ is

$$\begin{aligned}
E(X) &= \sum_{x \in E_X} x f_X(x) \\
&= \frac{1}{\binom{N}{n}} \sum_{x=\max(0,n-N+a)}^{\min(n,a)} x \binom{a}{x} \binom{N-a}{n-x} \\
&= \frac{a}{\binom{N}{n}} \sum_{x=\max(1,n-N+a)}^{\min(n,a)} \binom{a-1}{x-1} \binom{N-a}{n-x} \\
&= \frac{a}{\binom{N}{n}} \sum_{x=\max(0,n-N+a-1)}^{\min(n-1,a-1)} \binom{a-1}{x} \binom{(N-1)-(a-1)}{(n-1)-x} \\
&= \frac{a\binom{N-1}{n-1}}{\binom{N}{n}} \\
&= \frac{an}{N};
\end{aligned}$$

$$\begin{aligned}
E(X^2) &= \sum_{x \in E_X} x^2 f_X(x) \\
&= \frac{1}{\binom{N}{n}} \sum_{x=\max(0,n-N+a)}^{\min(n,a)} x^2 \binom{a}{x} \binom{N-a}{n-x} = \frac{1}{\binom{N}{n}} \sum_{x=\max(1,n-N+a)}^{\min(n,a)} x \frac{a!}{(a-x)!(x-1)!} \binom{N-a}{n-x} \\
&= \frac{1}{\binom{N}{n}} \sum_{x=\max(1,n-N+a)}^{\min(n,a)} (x-1+1) \frac{a!}{(a-x)!(x-1)!} \binom{N-a}{n-x} \\
&= \frac{1}{\binom{N}{n}} \left\{ a \sum_{x=\max(1,n-N+a)}^{\min(n,a)} \binom{a-1}{x-1} \binom{N-a}{n-x} + a(a-1) \sum_{x=\max(2,n-N+a)}^{\min(n,a)} \binom{a-2}{x-2} \binom{N-a}{n-x} \right\} \\
&= \frac{1}{\binom{N}{n}} \left\{ a \sum_{x=\max(0,n-N+a-1)}^{\min(n-1,a-1)} \binom{a-1}{x} \binom{(N-1)-(a-1)}{(n-1)-x} + a(a-1) \right. \\
&\quad \left. \sum_{x=\max(0,n-N+a-2)}^{\min(n-2,a-2)} \binom{a-2}{x} \binom{(N-2)-(a-2)}{(n-2)-x} \right\} \\
&= \frac{a\binom{N-1}{n-1}}{\binom{N}{n}} + \frac{a(a-1)\binom{N-2}{n-2}}{\binom{N}{n}} \\
&= \frac{an}{N} + \frac{a(a-1)n(n-1)}{N(N-1)};
\end{aligned}$$

$$Var(X) = E(X^2) - (E(X))^2 = \frac{an}{N} + \frac{a(a-1)n(n-1)}{N(N-1)} - \frac{a^2n^2}{N^2} = n\left(\frac{a}{N}\right)\left(1 - \frac{a}{N}\right)\left(\frac{N-n}{N-1}\right).$$

Example 2. *An urn contains* 6 *red balls and* 14 *black balls.* 5 *balls are drawn randomly without replacement. What is the probability that exactly* 4 *red balls are drawn?*

Solution: Let us label the drawing of a red as success and the drawing of a red as a failure. Let X be the number of red balls drawn. Then $X \sim \text{Hyp}(6, 5, 20)$. Hence, the required probability is $P(\{X = 4\}) = \frac{\binom{6}{4}\binom{14}{1}}{\binom{20}{5}}$.

2. Poisson Distribution

Suppose some event E is occurring randomly over a period of time. Let X be the number of times the event E has occurred in an unit interval (say $(0, 1]$).

Assumptions:

(1) For each infinitesimal subinterval $(\frac{k-1}{n}, \frac{k}{n}], k = 1, 2, \ldots, n$, the probability that the event E will occur in this subinterval is $\frac{\lambda}{n}$ and the probability that the event E will not occur in this subinterval is $1 - \frac{\lambda}{n}$, where $\lambda > 0$ is a given constant;
(2) chance of two or more occurrences of the event E in each infinitesimal subinterval $(\frac{k-1}{n}, \frac{k}{n}], k = 1, 2, \ldots, n$, is so small that it can be neglected;
(3) if $(\frac{j-1}{n}, \frac{j}{n}]$ and $(\frac{k-1}{n}, \frac{k}{n}](1 \leq j < k \leq n)$ are disjoint subintervals then the number of times the event E occurs in the interval $(\frac{j-1}{n}, \frac{j}{n}]$ is independent of the number of times the event E occurs in the interval $(\frac{k-1}{n}, \frac{k}{n}]$.

Remark 3. *Such type of events is known as rare events. It means that two such events are extremely unlikely to occur simultaneously or within a very short period of time. Arrivals of jobs, telephone calls, e-mail messages, traffic accidents, network blackouts, virus attacks, errors in software, floods, and earthquakes are examples of rare events.*

Under the above assumptions, in each infinitesimal subinterval $(\frac{k-1}{n}, \frac{k}{n}], k = 1, 2, \ldots, n$, event E can occur only 1 or 0 times and the probability of occurrence of event E in each of these subintervals is the same $(\frac{\lambda}{n})$. If we label the occurrence of event E in any of these subintervals as success and its non-occurrence as failure, then we have a sequence of n independent Bernoulli trials with probability of success in each trial as $p_n = \frac{\lambda}{n}$. Therefore, $X \equiv X_n \sim \text{Bin}(n, p_n)$, where $p_n = \frac{\lambda}{n}$. The p.m.f. of X is given by

$$\begin{aligned} f_n(k) &= \binom{n}{k} p_n^x (1 - p_n)^{n-k}, \text{ if } k = 1, 2, \ldots, n \\ &= \frac{1}{k!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right)(np_n)^k\left(1 - \frac{np_n}{n}\right)^n (1 - p_n)^{-k} \text{ if } k = 1, 2, \ldots, n \end{aligned}$$

Since $np_n = \lambda$ and $p_n \to 0$ as $n \to \infty$, $f_n(k) \to \frac{e^{-\lambda}\lambda^k}{k!}$ as $n \to \infty$.

Definition 4. *A discrete type random variable X is said to follow a Poisson distribution with parameter $\lambda > 0$ (written as $X \sim P(\lambda)$) if its support is $E_X = \{0, 1, 2, \ldots\}$ and its probability mass function is given by*

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

Remark 5. *From above discussion, it is clear that a Binomial distribution $Bin(n, p)$ with large n and small p can be approximated by a Poisson distribution $P(\lambda)$, where $\lambda = np$.*

Now, the m.g.f. of $X \sim \mathrm{P}(\lambda)$ is

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \sum_{x \in E_X} e^{tx} f_X(x) \\ &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{\lambda(e^t-1)}, \ t \in \mathbb{R} \end{aligned}$$

Therefore,

$$\begin{aligned} &M_X^{(1)}(t) = \lambda e^t e^{\lambda(e^t-1)}, \ t \in \mathbb{R}; \\ &M_X^{(2)}(t) = \lambda e^t e^{\lambda(e^t-1)} + (\lambda e^t)^2 e^{\lambda(e^t-1)}, \ t \in \mathbb{R}; \\ &E(X) = M_X^{(1)}(0) = \lambda; \\ &E(X^2) = M_X^{(2)}(0) = \lambda + \lambda^2; \\ &\text{and } Var(X) = E(X^2) - (E(X))^2 = \lambda. \end{aligned}$$

Example 6. *Ninety-seven percent of electronic messages are transmitted with no error. What is the probability that out of* 200 *messages, at least* 195 *will be transmitted correctly?*

Solution: Let us label the transmission of messages with no error as success and otherwise as failure. Let X be the number of correctly transmitted messages. Then $X \sim \mathrm{Bin}(200, 0.97)$. Hence the required probability is

$$P(X \geq 195) = 1 - P(X \leq 194) = 1 - \sum_{x=0}^{194} \binom{200}{x} (0.97)^x (0.03)^{200-x}.$$

X cannot be approximated by the Poisson distribution because success probability is too large.

Let Y be the number of failures. Then $Y \sim \mathrm{Bin}(200, 0.03)$. Then Y can be approximated by the Poisson distribution $Z \sim \mathrm{P}(6)$ (since $np = 200 \times 0.03 = 6$). Hence the required probability is

$$P(X \geq 195) = P(Y \leq 5) \approx P(Z \leq 5) = \sum_{x=0}^{5} \frac{e^{-6}6^x}{x!}.$$

Uniform and Normal Distribution

1. Uniform or Rectangular Distribution

Let α and β be two real numbers such that $-\infty < \alpha < \beta < \infty$. A continuous random variable X is said to have a uniform (or rectangular) distribution over the interval (α, β) (written as $X \sim U(\alpha, \beta)$) if probability density function of X is given by

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

Now, the r-th moment of $X \sim U(\alpha, \beta)$ is

$$\begin{aligned} E(X^r) &= \int_{-\infty}^{\infty} x^r f_X(x) dx \\ &= \int_{\alpha}^{\beta} \frac{x^r}{\beta - \alpha} dx \\ &= \frac{\beta^{r+1} - \alpha^{r+1}}{(r+1)(\beta - \alpha)} \\ &= \frac{\beta^r + \beta^{r-1}\alpha + \cdots + \beta\alpha^{r-1} + \alpha^r}{r+1} \end{aligned}$$

Hence

$$\begin{aligned} E(X) &= \frac{\alpha + \beta}{2}; \\ E(X^2) &= \frac{\beta^2 + \beta\alpha + \alpha^2}{3}; \\ Var(X) &= E(X^2) - (E(X))^2 = \frac{(\beta - \alpha)^2}{12}. \end{aligned}$$

The m.g.f. of $X \sim U(\alpha, \beta)$ is

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\ &= \int_{\alpha}^{\beta} \frac{e^{tx}}{\beta - \alpha} dx \\ &= \begin{cases} \frac{e^{t\beta} - e^{t\alpha}}{(\beta-\alpha)t}, & \text{if } t \neq 0 \\ 1, & \text{if } t = 0 \end{cases} . \end{aligned}$$

The d.f. of $X \sim U(\alpha, \beta)$ is

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)dt \\ &= \begin{cases} 0, \text{ if } x < \alpha \\ \frac{x-\alpha}{\beta-\alpha}, \text{ if } \alpha \le x < \beta \\ 1, \text{ if } x \ge \beta \end{cases} \end{aligned}$$

Remark 1. *Let $X \sim U(\alpha, \beta)$ and $Y = \frac{X-\alpha}{\beta-\alpha}$. Then the d.f. of Y is*

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(X \le \alpha + (\beta - \alpha)y) \\ &= \begin{cases} 0, \text{ if } \alpha + (\beta-\alpha)y < \alpha \\ \frac{\alpha+(\beta-\alpha)y-\alpha}{\beta-\alpha}, \text{ if } \alpha \le \alpha + (\beta-\alpha)y < \beta \\ 1, \text{ if } \alpha + (\beta-\alpha)y \ge \beta \end{cases} \\ &= \begin{cases} 0, \text{ if } y < 0 \\ y, \text{ if } 0 \le y < 1 \\ 1, \text{ if } y \ge 1 \end{cases} \end{aligned}$$

Clearly, F_Y is not differentiable at 0 and 1. Hence, the p.d.f. of Y is

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

Therefore, $Y \sim U(0, 1)$.

Example 2. *Let $a > 0$ be a real constant. A point X is chosen at random on the interval $(0, a)$ (i.e., $X \sim U(0, a)$).*

(1) If Y denotes the area of equilateral triangle having sides of length X, find the mean and variance of Y.
(2) If the point X divides the interval $(0, a)$ into subintervals $I_1 = (0, X)$ and $I_2 = [X, a)$, find the probability that the larger of these two subintervals is at least the double of the size of the smaller subinterval.

Solution:

(1) We have $Y = \frac{\sqrt{3}}{4}X^2$. Then

$$\begin{aligned} E(Y) &= \frac{\sqrt{3}}{4}E(X^2) = \frac{\sqrt{3}}{12}a^2; \\ E(Y^2) &= \frac{3}{16}E(X^4) = \frac{3}{80}a^4; \\ Var(Y) &= E(Y^2) - (E(Y))^2 = \frac{a^4}{80}. \end{aligned}$$

(2) The required probability is

$$\begin{aligned} p &= P(\{\max(X, a-X) \geq 2\min(X, a-X)\}) \\ &= P(\{a - X \geq 2X, X \leq \frac{a}{2}\}) + P(\{X \geq 2(a-X), X > \frac{a}{2}\}) \\ &= P(X \leq \frac{a}{3}\}) + P(\{X \geq \frac{2a}{3}\}) \\ &= F_X(\frac{a}{3}) + 1 - F_X(\frac{2a}{3}) \\ &= \frac{1}{3} + 1 - \frac{2}{3} = \frac{2}{3} \end{aligned}$$

2. Normal or Gaussian Distribution

(1) Let $\mu \in \mathbb{R}$ and $\sigma > 0$ be real constants. A continuous random variable X is said to have a normal (or Gaussian) distribution with parameters μ and σ^2 (written as $X \sim N(\mu, \sigma^2)$) if probability density function of X is given by

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

(2) The $N(0,1)$ distribution is called the standard normal distribution. The p.d.f. and the d.f. of $N(0,1)$ distributions will be denoted by ϕ and Φ respectively, i.e.,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \ -\infty < z < \infty$$

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx.$$

(3) We know that $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$ and $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$.

Clearly if $X \sim N(\mu, \sigma^2)$, then

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

Thus the distribution of X is symmetric about μ. Hence,

$$X \sim N(\mu, \sigma^2) \Rightarrow F_X(\mu - x) + F_X(\mu + x) = 1, \ \forall \ x \in \mathbb{R} \text{ and } F_X(\mu) = \frac{1}{2}.$$

In particular,

$$\boxed{\Phi(-z) = 1 - \Phi(z), \ \forall \ z \in \mathbb{R} \text{ and } \Phi(0) = \frac{1}{2}.}$$

Suppose that $X \sim N(\mu, \sigma^2)$. Then the p.d.f. of $Z = \frac{X-\mu}{\sigma}$ is given by

$$\begin{aligned} f_Z(z) &= f_X(\mu + \sigma z)|\sigma|, \ -\infty < z < \infty \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \ -\infty < z < \infty \end{aligned}$$

i.e.,

$$\boxed{X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0,1).}$$

Thus

$$\boxed{X \sim N(\mu, \sigma^2) \Rightarrow F_X(x) = P(\{X \leq x\}) = P\left(\left\{Z \leq \frac{x-\mu}{\sigma}\right\}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right), \ \forall \ x \in \mathbb{R}.}$$

Now, the m.g.f. of $X \sim N(\mu, \sigma^2)$ is

$$\begin{aligned}
M_X(t) &= E(e^{tX}) \\
&= \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \frac{e^{\mu t}}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2+\sqrt{2}\sigma t y} dy \text{ (by putting } \frac{x-\mu}{\sqrt{2}\sigma} = y) \\
&= \frac{e^{(\mu t + \frac{\sigma^2 t^2}{2})}}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-(y-\frac{\sqrt{2}\sigma t}{2})^2} dy \\
&= e^{(\mu t + \frac{\sigma^2 t^2}{2})}, \ \forall\ t \in \mathbb{R}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&M_X^{(1)}(t) = (\mu + \sigma^2 t) e^{(\mu t + \frac{\sigma^2 t^2}{2})}, \ \forall\ t \in \mathbb{R}; \\
&M_X^{(2)}(t) = (\sigma^2 + (\mu + \sigma^2 t)^2) e^{(\mu t + \frac{\sigma^2 t^2}{2})}, \ \forall\ t \in \mathbb{R}; \\
&E(X) = M_X^{(1)}(0) = \mu; \\
&E(X^2) = M_X^{(2)}(0) = \mu^2 + \sigma^2; \\
&\text{and } Var(X) = E(X^2) - (E(X))^2 = \sigma^2.
\end{aligned}$$

Gamma and Exponential Distribution

Consider the improper integral $\int_0^\infty e^{-t}t^{\alpha-1}dt = \int_0^1 e^{-t}t^{\alpha-1}dt + \int_1^\infty e^{-t}t^{\alpha-1}dt$, where $\alpha \in \mathbb{R}$. By Limit comparison test, $\int_0^1 e^{-t}t^{\alpha-1}dt$ converges, for all $\alpha > 0$ and the $\int_1^\infty e^{-t}t^{\alpha-1}dt$ converges, for all $\alpha \in \mathbb{R}$. Hence, the $\int_0^\infty e^{-t}t^{\alpha-1}dt$ is convergent if and only if $\alpha > 0$.

Definition 1. *The function* $\Gamma : (0,\infty) \longrightarrow (0,\infty)$*, defined by,*

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

is called the gamma function.

Properties:

(1) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\alpha > 0$.
(2) $\Gamma(n) = (n-1)!$, $n \in \mathbb{N}$ with the convention that $0! = 1$.
(3) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. In general, for $n \in \mathbb{N} \cup \{0\}$, we have $\Gamma(\frac{2n+1}{2}) = \frac{(2n)!}{n!4^n}\sqrt{\pi}$.

1. Gamma Distribution

A continuous random variable X is said to have a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$ (written as $X \sim G(\alpha,\lambda)$) if probability density function of X is given by

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

Now, the r-th moment of $X \sim G(\alpha,\lambda)$ is

$$\begin{aligned} E(X^r) &= \int_{-\infty}^\infty x^r f_X(x)dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^r e^{-\lambda x}x^{\alpha-1}dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty e^{-\lambda x}x^{(\alpha+r)-1}dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)\lambda^{(\alpha+r)}}\int_0^\infty e^{-t}t^{(\alpha+r)-1}dx, \text{ (by putting } \lambda x = t) \\ &= \frac{\Gamma(\alpha+r)}{\Gamma(\alpha)\lambda^r} \\ &= \frac{\alpha(\alpha+1)\cdots(\alpha+r-1)}{\lambda^r} \end{aligned}$$

Hence

$$E(X) = \frac{\alpha}{\lambda};$$
$$E(X^2) = \frac{\alpha(\alpha+1)}{\lambda^2};$$
$$Var(X) = E(X^2) - (E(X))^2 = \frac{\alpha}{\lambda^2}.$$

The m.g.f. of $X \sim G(\alpha, \lambda)$ is

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \int_{-\infty}^{\infty} e^{tx} f_X(x)dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{\infty} e^{-(\lambda-t)x} x^{\alpha-1} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)(\lambda-t)^\alpha} \int_0^{\infty} e^{-z} z^{\alpha-1} dz, \text{ if } t < \lambda \text{ (by putting } \lambda - t = z) \\ &= \left(\frac{\lambda}{\lambda-t}\right)^\alpha, \text{ if } t < \lambda. \end{aligned}$$

Remark 2. *Let $X \sim G(\alpha, \lambda)$ and $h : \mathbb{R} \longrightarrow \mathbb{R}$ be a function defined by $h(x) = \lambda x$. Since the support of X is $E_X = (0, \infty)$, the support of $Z = h(X) = \lambda X$ is $E_Z = (0, \infty)$. Clearly, h is strictly increasing on E_X. Therefore, the p.d.f. of $Z = \lambda X$ is*

$$\begin{aligned} f_Z(z) &= \begin{cases} f_X(h^{-1}(z)) \left|\frac{d}{dz} h^{-1}(z)\right|, & \text{if } z > 0 \\ 0, & \text{if } z \leq 0 \end{cases} \\ &= \begin{cases} \frac{e^{-z} z^{\alpha-1}}{\Gamma(\alpha)}, & \text{if } z > 0 \\ 0, & \text{if } z \leq 0 \end{cases} \end{aligned}$$

Hence $Z \sim G(\alpha, 1)$.

2. Exponential Distribution

A $G(1, \lambda)$ distribution is called an exponential distribution with parameter $\lambda > 0$ and it is denoted by $\text{Exp}(\lambda)$. Thus p.d.f. of $\text{Exp}(\lambda)$ is

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

If $X \sim \text{Exp}(\lambda)$, then

$$\begin{aligned} E(X^r) &= \frac{r!)}{\lambda^r}; \\ E(X) &= \frac{1}{\lambda}; \\ E(X^2) &= \frac{2}{\lambda^2}; \\ Var(X) &= E(X^2) - (E(X))^2 = \frac{1}{\lambda^2}; \\ M_X(t) &= \frac{\lambda}{\lambda - t}, \text{ if } t < \lambda. \end{aligned}$$

The d.f. of $X \sim \text{Exp}(\lambda)$ is

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)dt \\ &= \begin{cases} \int_0^x \lambda e^{-\lambda t}, \text{ if } x > 0 \\ 0, \text{ if } x \leq 0 \end{cases} \\ &= \begin{cases} 1 - e^{-\lambda x}, \text{ if } x > 0 \\ 0, \text{ if } x \leq 0 \end{cases} \end{aligned}$$

Remark 3. (1) *A Poisson Process is a model for a series of discrete event where the average time between events is known, but the exact timing of events is random.*

(2) *The exponential distribution occurs naturally if we consider the distribution of the length of intervals between successive events in a Poisson process or, equivalently, the distribution of the interval (i.e. the waiting time) before the first event.*

Example 4. *The waiting time for occurrence of an event E (say repair time of a machine) is exponentially distributed with mean of 30 minutes. Find the conditional probability that the waiting time for occurrence of event E is at least 5 hours given that it has not occurred in the first 3 hours.*

Solution: Let X be the waiting time (in hours) for the occurrence of event E. Then $X \sim \text{Exp}(2)$. Hence, the required probability is $P(\{X > 5\}|\{X > 3\}) = \frac{P(\{X>5\})}{P(\{X>3\})} = \frac{e^{-10}}{e^{-6}} = e^{-4}$.

Random Vector

Let $(\mathcal{S}, \Sigma, P)$ be a probability space. A (univariate) random variable describes a numerical quantity of a typical outcome of a random experiment. In many experiments an observation is expressed as a family of several separate numerical quantities and we may be interested in simultaneously studying all of them together. Consider the following example.

Example 1. *Two distinguishable dice (labelled as D_1 and D_2) are thrown simultaneously. The sample space is $\mathcal{S} = \{(i,j) : i,j \in \{1,2,\ldots,6\}\}$. For $(i,j) \in \mathcal{S}$ define*

$$X_1((i,j)) = i + j = \text{ sum of number of dots on uppermost faces of two dice}$$

and

$X_2((i,j)) = |i-j| =$ *absolute difference of number of dots on uppermost faces of two dice.*

It may be of interest to study numerical characteristics X_1 and X_2 simultaneously. These considerations lead to the study of the function $\underline{X} = (X_1, X_2) : \mathcal{S} \to \mathbb{R}$

Notations.

- We denote by $\mathbb{R}^n$ the n-dimensional Euclidean space, i.e.,
$$\mathbb{R}^n = \{\underline{x} = (x_1, x_2, \ldots, x_n) : x_i \in \mathbb{R},\ i = 1, 2, \ldots, n\}.$$
- For $i = 1, 2, \ldots, n$, let $X_i : \mathcal{S} \to \mathbb{R}$ be any functions. Then the function $\underline{X} = (X_1, X_2, \ldots, X_n) : \mathcal{S} \to \mathbb{R}^n$ is defined as
$$\underline{X}(w) = (X_1(w), X_2(w), \ldots, X_n(w)),\ w \in \mathcal{S}.$$
- For $A \subseteq \mathbb{R}^n$,
$$\underline{X}^{-1}(A) = \{w \in \mathcal{S} : \underline{X}(w) \in A\}.$$
- For $\underline{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, we denote by $(-\underline{\infty}, \underline{x}]$ the n-dimensional interval
$$(-\underline{\infty}, \underline{x}] = (-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_n].$$

Definition 2. *A function $\underline{X} : \mathcal{S} \longrightarrow \mathbb{R}^n$ is called an $n-$dimensional random vector (RV) if $\underline{X}^{-1}((-\underline{\infty}, \underline{x}]) \in \Sigma$, for all $\underline{x} \in \mathbb{R}^n$. That is, $\{w \in \mathcal{S} : X_1(w) \leq x_1, X_2(w) \leq x_2, \ldots, X_n(w) \leq x_n\} \in \Sigma$.*

Example 3. *Let $A, B \subseteq \mathcal{S}$. Define $\underline{X} = (X_1, X_2) : \mathcal{S} \to \mathbb{R}^2$ by*

$$X_1(w) = I_A(w) = \begin{cases} 1, \text{if } w \in A, \\ 0, \text{if } w \notin A; \end{cases}$$

and

$$X_2(w) = I_B(w) = \begin{cases} 1, \text{if } w \in B, \\ 0, \text{if } w \notin B. \end{cases}$$

Then $\underline{X}$ is an RV if and only if A and B are events. (Prove!)

Theorem 4. *Let $\underline{X} = (X_1, X_2, \ldots, X_n) : \mathcal{S} \to \mathbb{R}^n$ be a given function. Then $\underline{X}$ is a random vector if and only if $X_1, X_2, \ldots, X_n$ are random variables.*

Proof. Exercise. □

Remark 5. *If $\mathcal{S}$ is finite or countable and $\Sigma = \mathcal{P}(\Sigma)$, then any function $\underline{X} = (X_1, X_2, \ldots, X_n) : \mathcal{S} \to \mathbb{R}^n$ is a random vector.*

Joint Cumulative Distribution Function

Definition 6. *Let $\underline{X} = (X_1, X_2, \ldots, X_n) : \mathcal{S} \to \mathbb{R}^n$ be a random vector. The function $F_{\underline{X}} : \mathbb{R}^n \to \mathbb{R}$, defined by,*

$$F_{\underline{X}}(x_1, x_2, \ldots, x_n) = P(\{w \in \mathcal{S} : X_1(w) \leq x_1, X_2(w) \leq x_2, \ldots, X_n(w) \leq x_n\}),\ \forall\ \underline{x} \in \mathbb{R}^n,$$

is called the ***joint cumulative distribution function*** *(joint c.d.f) or the* ***joint distribution function*** *(d.f) of the random vector $\underline{X}$.*

The joint distribution function of any subset of random variables $X_1, X_2, \ldots, X_n$ is called a marginal distribution function of $F_{\underline{X}}$.

Remark 7.
(1) *As in the case of random variables, the set $\{w \in \mathcal{S} : X_1(w) \leq x_1, X_2(w) \leq x_2, \ldots, X_n(w) \leq x_n\}$ will be denoted by $\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}$.*
(2) *In this course, we will mainly study 2- (and sometimes 3-) dimensional random vectors.*
(3) *Let $\underline{X} = (X, Y) : \mathcal{S} \to \mathbb{R}^2$ be a random vector. The joint c.d.f. is a map $F_{\underline{X}} : \mathbb{R}^2 \to \mathbb{R}$, defined by,*

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

(4) *The c.d.f. of X and Y are called a marginal c.d.f. of $F_{\underline{X}}$.*

Proposition 8. *Let $\underline{X} = (X, Y) : \mathcal{S} \to \mathbb{R}^2$ be a random vector with joint c.d.f. $F_{\underline{X}}$. Then the marginal c.d.f. of X and Y are given by*

$$F_X(x) = \lim_{y \to \infty} F_{\underline{X}}(x, y) \text{ and } F_Y(y) = \lim_{x \to \infty} F_{\underline{X}}(x, y)$$

Remark 9. *Let $(a_1, b_1), (a_2, b_2) \in \mathbb{R}^2$. Then we know that*

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a).$$

Now,

$$\begin{aligned}
&P(a_1 < X \leq b_1, a_2 < Y \leq b_2) \\
&= P(a_1 < X \leq b_1, Y \leq b_2) - P(a_1 < X \leq b_1, Y \leq a_2) \\
&= [P(X \leq b_1, Y \leq b_2) - P(X \leq a_1, Y \leq b_2)] \\
&\qquad - [P(X \leq b_1, Y \leq a_2) - P(X \leq a_1, Y \leq a_2)] \\
&= F_{\underline{X}}(b_1, b_2) - F_{\underline{X}}(a_1, b_2) - F_{\underline{X}}(b_1, a_2) + F_{\underline{X}}(a_1, a_2).
\end{aligned}$$

Theorem 10. *Let $F_{\underline{X}}$ be the joint cumulative distribution function of a random vector $\underline{X} = (X, Y)$. Then*

(1) $\lim\limits_{\substack{x \to \infty \\ y \to \infty}} F_{\underline{X}}(x, y) = 1.$
(2) $\lim_{y \to -\infty} F_{\underline{X}}(x, y) = 0$ *and* $\lim_{x \to -\infty} F_{\underline{X}}(x, y) = 0.$
(3) *$F_{\underline{X}}(x, y)$ is right continuous and nondecreasing in each argument (keeping other argument fixed).*
(4) *For each $(a_1, b_1] \times (a_2, b_2]$ in $\mathbb{R}^2$,*

$$\Delta = F_{\underline{X}}(b_1, b_2) - F_{\underline{X}}(a_1, b_2) - F_{\underline{X}}(b_1, a_2) + F_{\underline{X}}(a_1, a_2) \geq 0.$$

Theorem 11. *Let $G : \mathbb{R}^2 \to \mathbb{R}$ be a function which satisfies properties (1) − (4) of Theorem 10. Then there exists a probability space $(\mathcal{S}, \Sigma, P)$ and a random vector $\underline{X} = (X_1, X_2, \ldots, X_n)$ defined on $(\mathcal{S}, \Sigma, P)$ such that G is the distribution function of $\underline{X}$.*

Example 12. *Let $G : \mathbb{R}^2 \to \mathbb{R}$ be defined by*

$$G(x,y) = \begin{cases} x, \textit{if } 0 \le x < 1,\ y \ge 1, \\ y^2, \textit{if } x \ge 1,\ 0 \le y < 1, \\ 1, \textit{if } x \ge 1,\ y \ge 1, \\ 0, \textit{otherwise.} \end{cases}$$

Show that G is not a distribution function of any random vector (X, Y).

Solution. Clearly G satisfies properties $(1) - (3)$ of Theorem 10.

For $(a_1, b_1] \times (a_2, b_2]$, where $a_1, a_2 \in [0, 1)$, $b_1, b_2 \in [1, \infty)$ and $a_1 + a_2^2 > 1$. Then

$$G(b_1, b_2) - G(a_1, b_2) - G(b_1, a_2) + G(a_1, a_2) = 1 - a_1 - a_2^2 + 0 < 0.$$

Thus, G is not a joint c.d.f. of any random vector.

Example 13. *Consider the function $G : \mathbb{R}^2 \to \mathbb{R}$ defined by*

$$G(x,y) = \begin{cases} xy^2, \textit{if } 0 \le x < 1,\ 0 \le y < 1, \\ x, \textit{if } 0 \le x < 1,\ y \ge 1, \\ y^2, \textit{if } x \ge 1,\ 0 \le y < 1, \\ 1, \textit{if } x \ge 1,\ y \ge 1, \\ 0, \textit{otherwise.} \end{cases}$$

(1) *Show that G is a joint c.d.f. of some random vector (X, Y).*
(2) *Find the marginal c.d.f. of X and Y.*

Solution. Clearly $\lim\limits_{\substack{x\to\infty \\ y\to\infty}} G(x,y) = 1$. For fixed $x \in \mathbb{R}$, $\lim_{y\to-\infty} G(x,y) = 0$ and for fixed $y \in \mathbb{R}$, $\lim_{x\to-\infty} G_(x,y) = 0$.

We note that if $y < 0$, then $G(x,y) = 0$ for all $x \in \mathbb{R}$. Moreover,

$$G(x,y) = \begin{cases} 0, \text{if } x < 0, \\ xy^2, \text{if } 0 \le x < 1,\ 0 \le y < 1, \\ y^2, \text{if } x \ge 1, \end{cases}$$

and

$$G(x,y) = \begin{cases} 0, \text{if } x < 0, \\ x, \text{if } 0 \le x < 1,\ y \ge 1, \\ 1, \text{if } x \ge 1. \end{cases}$$

One can see that for $y \in \mathbb{R}$, $G(x,y)$ is a continuous (and hence right continuous) function of x. Similarly, for each $x \in \mathbb{R}$, $G(x,y)$ is a continuous function of y

Furthermore, $G(x,y)$ is non-decreasing in each argument keeping other argument fixed.

For $(a_1, b_1] \times (a_2, b_2]$, we need to show that $\Delta = G(b_1, b_2) - G(a_1, b_2) - G(b_1, a_2) + G(a_1, a_2) \ge 0$. We consider the following cases.

(1) $a_1 < 0$. Then $\Delta = G(b_1, b_2) - G(b_1, a_2) \ge 0$ as G is nondecreasing.
(2) $a_2 < 0$.
(3) $0 \le a_1 < 1$, $0 \le a_2 < 1$, $0 \le b_1 < 1$, $0 \le b_2 < 1$.
(4) $0 \le a_1 < 1$, $0 \le a_2 < 1$, $0 \le b_1 < 1$, $b_2 \ge 1$.
(5) $0 \le a_1 < 1$, $0 \le a_2 < 1$, $b_1 \ge 1$, $0 \le b_2 < 1$.
(6) $0 \le a_1 < 1$, $0 \le a_2 < 1$, $b_1 \ge 1$, $b_2 \ge 1$.

(7) $0 \le a_1 < 1$, $a_2 \ge 1$, $0 \le b_1 < 1$, $b_2 \ge 1$.
(8) $0 \le a_1 < 1$, $a_2 \ge 1$, $b_1 \ge 1$, $b_2 \ge 1$.
(9) $a_1 \ge 1$, $0 \le a_2 < 1$, $b_1 \ge 1$, $0 \le b_2 < 1$.
(10) $a_1 \ge 1$, $0 \le a_2 < 1$, $b_1 \ge 1$, $b_2 \ge 1$.
(11) $a_1 \ge 1$, $a_2 \ge 1$, $b_1 \ge 1$, $b_2 \ge 1$.

In all these cases verify that $\Delta \ge 0$.

Therefore, $G(x, y)$ is a distribution function of some random vector (X, Y).

The marginal c.d.f. of X and Y are respectively

$$F_X(x) = \lim_{y \to \infty} G(x, y) = \begin{cases} 0, \text{if } x < 0, \\ x, \text{if } 0 \le x < 1, \\ 1, \text{if } x \ge 1, \end{cases}$$

and

$$F_Y(y) = \lim_{x \to \infty} G(x, y) = \begin{cases} 0, \text{if } y < 0, \\ y^2, \text{if } 0 \le y < 1, \\ 1, \text{if } y \ge 1. \end{cases}$$

Types of Random Vector

Let $(\mathcal{S}, \Sigma, P)$ be a probability space and let $\underline{X} = (X, Y) : \mathcal{S} \longrightarrow \mathbb{R}^2$ be a random vector with joint distribution function $F_{\underline{X}}$.

Notations.

- Let $\mathbb{B}_{\mathbb{R}^n}$ denote the set which contains all rectangles (Cartesian product of open, closed and semi-closed intervals) and their countable union and intersection.
- Let I_n be a rectangle in $\mathbb{R}^n$. We will denote by $\mathbb{B}_{I_n}$ the set which contains all rectangles contained in I_n and their countable union and intersection.

Definition 1. *$\underline{X}$ is a said to be a random vector of discrete type if there exists a nonempty finite or countable set $E_{\underline{X}} \subset \mathbb{R}^2$ such that $P(\underline{X} = \underline{x}) > 0$, for every $\underline{x} \in E_{\underline{X}}$, and $P(\underline{X} \in E_{\underline{X}}) = 1$.*

The set $E_{\underline{X}}$ is called the support of $\underline{X}$.

The function $f_{\underline{X}} : \mathbb{R}^2 \to \mathbb{R}$ defined by

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

is called the joint probability mass function of $\underline{X}$.

Remark 2. *Let $\underline{X}$ be a random vector of discrete type with support $E_{\underline{X}}$, joint d.f. $F_{\underline{X}}$ and joint p.m.f. $f_{\underline{X}}$.*

(1) $\sum_{\underline{x} \in E_{\underline{X}}} f_{\underline{X}}(\underline{x}) = 1$. *Moreover, $P(\underline{X} \in E^c_{\underline{X}}) = 0$ and $f_{\underline{X}}(\underline{x}) = 0$, $\forall\ \underline{x} \in E^c_{\underline{X}}$.*
(2) *For any $A \in \mathbb{B}_{\mathbb{R}^2}$,*

$$P(\underline{X} \in A) = \sum_{\underline{x} \in A \cap E_{\underline{X}}} f_{\underline{X}}(\underline{x}) = \sum_{\underline{x} \in E_{\underline{X}}} f_{\underline{X}}(\underline{x}) I_A(\underline{x}).$$

(3) *For $\underline{x} \in \mathbb{R}^2$,*

$$F_{\underline{X}}(\underline{x}) = P(\underline{X} \in (-\underline{\infty}, \underline{x}]) = \sum_{\underline{x} \in (-\underline{\infty}, \underline{x}] \cap E_{\underline{X}}} f_{\underline{X}}(\underline{x}).$$

Definition 3. *$\underline{X}$ is a said to be a random vector of continuous type if there exists a nonnegative function $f_{\underline{X}} : \mathbb{R}^2 \to \mathbb{R}$ such that*

$$F_{\underline{X}}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{\underline{X}}(x, y) dy dx.$$

The set $E_{\underline{X}} = \{\underline{x} \in \mathbb{R}^2 : f_{\underline{X}}(\underline{x}) > 0\}$ is called the support of $\underline{X}$.

The function $f_{\underline{X}}$ is called the joint probability density function of $\underline{X}$.

Remark 4. *Let $\underline{X}$ be a random vector of continuous type with support $E_{\underline{X}}$, joint d.f. $F_{\underline{X}}$ and joint p.d.f. $f_{\underline{X}}$.*

(1) *For any $\underline{x} \in \mathbb{R}^2$, $f_{\underline{X}}(\underline{x}) \geq 0$, and*

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\underline{X}}(x, y) dy dx = 1.$$

(2) *For any $\underline{x} \in \mathbb{R}^2$, $P(\underline{X} = \underline{x}) = 0$. Consequently, for any countable set $S \subset \mathbb{R}^2$, $P(\underline{X} \in S) = 0$.*

(3) *Let* $\underline{a} = (a_1, a_2)$, $\underline{b} = (b_1, b_2) \in \mathbb{R}^2$ *such that* $a_i < b_i$, $i = 1, 2$. *Let* $(\underline{a}, \underline{b}] = (a_1, a_2] \times (b_1, b_2]$. *Then*

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

Theorem 5. *Let* $\underline{X} = (X, Y) : \mathcal{S} \longrightarrow \mathbb{R}^2$ *be a random vector with joint distribution function* $F_{\underline{X}}$.

(1) *Suppose that* $\underline{X}$ *is of discrete type with support* $E_{\underline{X}}$ *and joint p.m.f.* $f_{\underline{X}}$. *Define*

$$R_x = \{y \in \mathbb{R} : (x, y) \in E_{\underline{X}}\}, \ R_y = \{x \in \mathbb{R} : (x, y) \in E_{\underline{X}}\}.$$

Then X *and* Y *are of discrete type with support*

$$E_X = \{x \in \mathbb{R} : (x, y) \in E_{\underline{X}} \text{ for some } y \in \mathbb{R}\}$$

and

$$E_Y = \{y \in \mathbb{R} : (x, y) \in E_{\underline{X}} \text{ for some } x \in \mathbb{R}\}$$

respectively. The marginal p.m.f.s of X *and* Y *are respectively given by*

$$f_X(x) = \begin{cases} \sum_{y \in R_x} f_{\underline{X}}(x, y), & \text{if } x \in E_X, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$f_Y(y) = \begin{cases} \sum_{x \in R_y} f_{\underline{X}}(x, y), & \text{if } y \in E_Y, \\ 0, & \text{otherwise}. \end{cases}$$

(2) *Suppose that* $\underline{X}$ *is of continuous type with support* $E_{\underline{X}}$ *and joint p.d.f.* $f_{\underline{X}}$. *Then* X *and* Y *are of continuous type with marginal p.d.f.s given by*

$$f_X(x) = \int_{-\infty}^{\infty} f_{\underline{X}}(x, y) dy \text{ and } f_Y(y) = \int_{-\infty}^{\infty} f_{\underline{X}}(x, y) dx$$

respectively.

Example 6. *Let* $\underline{X} = (X, Y)$ *be a random vector with joint p.m.f.*

$$f_{\underline{X}}(x, y) = \begin{cases} cy, & \text{if } (x, y) \in A, \\ 0, & \text{otherwise}; \end{cases}$$

where $A = \{(a, b) : a, b \in \{1, 2, \ldots, n\}, \ a \leq b\}$, $n \geq 2$ *is a fixed integer and* c *is a constant.*

(1) *Find the value of* c.
(2) *Find the marginal p.m.f.s of* X *and* Y.
(3) *Find* $P(X > Y)$, $P(X = Y)$ *and* $P(X < Y)$.

Solution.

(1) Clearly $c > 0$. The support $E_{\underline{X}}$ is A. Therefore, $\sum_{(x,y) \in E_{\underline{X}}} f_{\underline{X}}(x, y) = 1$. This implies that $c \sum_{y=1}^{n} \sum_{x=1}^{y} y = 1$ or $c \sum_{y=1}^{n} y^2 = 1$. Thus, $c = \frac{6}{n(n+1)(2n+1)}$.
(2) The support of X is $E_X = \{1, 2, \ldots, n\}$ and the support of Y is $E_Y = \{1, 2, \ldots, n\}$. For $x \in E_X$, we have $R_x = \{x, x + 1, \ldots, n\}$ and

$$\sum_{y \in R_x} f_{\underline{X}}(x, y) = c \sum_{y=x}^{n} y = c \left[\frac{n(n+1)}{2} - \frac{(x-1)x}{2} \right].$$

The marginal p.m.f. of X is then

$$f_X(x) = \begin{cases} \frac{3[n(n+1)-(x-1)x]}{n(n+1)(2n+1)}, & \text{if } x \in E_X, \\ 0, & \text{otherise.} \end{cases}$$

For $y \in E_Y$, we have $R_y = \{1, 2, \ldots, y\}$ and

$$\sum_{x \in R_y} f_{\underline{X}}(x, y) = c \sum_{x=1}^{y} y = cy^2.$$

The marginal p.m.f. of Y is then

$$f_Y(y) = \begin{cases} \frac{3y^2}{n(n+1)(2n+1)}, & \text{if } x \in E_Y, \\ 0, & \text{otherise.} \end{cases}$$

(3) Let $A = \{(a, b) : a > b\}$ and $B = \{(a, b) : a = b\}$. Then

$$\begin{aligned} P(X > Y) &= P(\underline{X} \in A) \\ &= \sum_{(x,y) \in E_{\underline{X}} \cap A} f_{\underline{X}}(x, y) \\ &= 0. \end{aligned}$$

$$\begin{aligned} P(X = Y) &= P(\underline{X} \in B) \\ &= \sum_{(x,y) \in E_{\underline{X}} \cap B} f_{\underline{X}}(x, y) \\ &= c \sum_{y=1}^{n} y = \frac{3}{2n+1}. \end{aligned}$$

Therefore, $P(X < Y) = \frac{2(n-1)}{2n+1}$.

Example 7. *Let $\underline{X} = (X, Y)$ be a random vector with joint p.d.f.*

$$f_{\underline{X}}(x, y) = \begin{cases} \frac{c}{x}, & \text{if } 0 < y < x < 1, \ c \in \mathbb{R}, \\ 0, & \text{otherwise.} \end{cases}$$

(1) *Find the value of c.*
(2) *Find the marginal p.d.f.s of X and Y.*
(3) *Find $P(X > 2Y)$.*

Solution.

(1) Since $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\underline{X}}(x, y) dx dy = 1$. This implies that $c \int_0^1 \int_0^x \frac{1}{x} dy dx = 1$ or $c \int_0^1 dx = 1$ or $c = 1$.
(2) The marginal p.d.f. of X is given by

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{\underline{X}}(x, y) dy \\ &= \begin{cases} \int_0^x \frac{1}{x} dy, & \text{if } 0 < x < 1, \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} 1, & \text{if } 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

The marginal p.d.f. of Y is given by

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{\underline{X}}(x,y)dx \\ &= \begin{cases} \int_y^1 \frac{1}{x}dx, & \text{if } 0<y<1, \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} -\ln y, & \text{if } 0<y<1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

(3) Let $A = \{(x,y) : x > 2y\}$. Then

$$\begin{aligned} P(X > 2Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\underline{X}}(x,y)I_A(x,y)dydx \\ &= \iint\limits_{0<2y<x<1} \frac{1}{x}dydx \\ &= \int_0^1 \int_0^{x/2} \frac{1}{x}dydx \\ &= \frac{1}{2}. \end{aligned}$$

Conditional Distributions and Independent random variables

1. Conditional Distributions

Definition 1. *Let $\underline{Z} = (X, Y)$ be a random vector of discrete type with support $E_{\underline{Z}}$, joint d.f. $F_{\underline{Z}}$ and joint p.m.f. $f_{\underline{Z}}$. Then X and Y are discrete type random variables.*

For a fixed y with $P(Y = y) > 0$, the function $f_{X|Y}(.|y) : \mathbb{R} \longrightarrow \mathbb{R}$ defined as

$$f_{X|Y}(x|y) = P(X = x|Y = y),\ \forall\ x \in \mathbb{R},$$

is called the conditional probability mass function of X, given $Y = y$. Thus, the conditional probability mass function of X, given $Y = y$, is

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{\underline{Z}}(x, y)}{f_Y(y)}$$
$$= \begin{cases} \frac{f_{\underline{Z}}(x,y)}{f_Y(y)},\ \text{if } x \in E_{X|Y=y} \\ 0,\ \text{otherwise}, \end{cases}$$

where $E_{X|Y=y} = \{x \in \mathbb{R} \mid (x, y) \in E_{\underline{Z}}\}$ and f_Y is the marginal p.m.f. of Y.

The conditional cumulative distribution function of X, given $Y = y$, is defined as

$$\begin{aligned} F_{X|Y}(x|y) &= P(X \le x|Y = y) \\ &= \frac{P(X \le x, Y = y)}{P(Y = y)} \\ &= \sum_{x_i \in E_{X|Y=y} \cap (-\infty, x]} \frac{f_{\underline{Z}}(x_i, y)}{f_Y(y)} \\ &= \sum_{x_i \le x} f_{X|Y}(x_i|y),\ \text{where } x_i \in E_{X|Y=y}. \end{aligned}$$

In the similar manner, we can define the conditional probability mass function and conditional cumulative distribution function of Y, given $X = x$, provided $P(X = x) > 0$.

Definition 2. *Let $\underline{Z} = (X, Y)$ be a random vector of continuous type with joint c.d.f. $F_{\underline{Z}}$ and joint p.d.f. $f_{\underline{Z}}$. Then X and Y are continuous type random variables. Let $y \in \mathbb{R}$ be such that $f_Y(y) > 0$, where $f_Y(y) > 0$ is the marginal p.d.f. of Y.*

The function $f_{X|Y}(.|y) : \mathbb{R} \longrightarrow \mathbb{R}$ defined as

$$f_{X|Y}(x|y) = \frac{f_{\underline{Z}}(x, y)}{f_Y(y)},\ \forall\ x \in \mathbb{R},$$

is called the conditional probability density function of X, given $Y = y$.

Also, the conditional cumulative distribution function of X, given $Y = y$, is defined as

$$\begin{aligned} F_{X|Y}(x|y) &= \int_{-\infty}^{x} f_{X|Y}(t|y)dt \\ &= \int_{-\infty}^{x} \frac{f_{\underline{Z}}(t, y)}{f_Y(y)}dt \end{aligned}$$

In the similar manner, we can define the conditional probability density function and conditional cumulative distribution function of Y, given $\{X = x\}$, provided $f_X(x) > 0$, where $f_X(x) > 0$ is the marginal p.d.f. of X.

Note: Definition 1 and 2 can be generalized if we replace random variables X and Y by random vectors $\underline{X}$ and $\underline{Y}$.

Example 3. *Let $\underline{Z} = (X, Y)$ be a random vector with joint p.d.f.*

$$f(x,y) = \begin{cases} 6xy(2-x-y), & \text{if } 0 < x < 1, 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Then find the conditional p.d.f. of X, given $Y = y$, where $0 < y < 1$.

Solution: The conditional p.d.f. of X, given $Y = y$, is

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x,y)}{f_Y(y)} \\ &= \begin{cases} \frac{6xy(2-x-y)}{\int_0^1 6xy(2-x-y)dx}, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{6x(2-x-y)}{4-3y}, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Example 4. *Let $\underline{Z} = (X, Y, Z)$ be a random vector with joint p.m.f.*

$$f(x,y,z) = \begin{cases} \frac{xyz}{72}, & \text{if } (x,y,z) \in \{1,2\} \times \{1,2,3\} \times \{1,3\} \\ 0, & \text{otherwise} \end{cases}$$

(1) *Find the conditional p.m.f. of X, given $(Y, Z) = (2, 1)$.*
(2) *Find the conditional p.m.f. of (X, Z), given $Y = 3$.*

Solution:

(1) The conditional p.m.f. of X, given $(Y, Z) = (2, 1)$, is

$$\begin{aligned} f_{X|(Y,Z)}(x|(2,1)) &= \frac{f(x,2,1)}{P((Y,Z)=(2,1))} \\ &= \begin{cases} \frac{2x}{72P(Y=2,Z=1)}, & \text{if } x \in E_{X|(Y,Z)=(2,1)} = \{x \in \mathbb{R} \mid (x,2,1) \in E_{\underline{Z}}\} \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{2x}{72P(Y=2,Z=1)}, & \text{if } x \in \{1,2\} \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Now, $P(Y = 2, Z = 1) = \sum_{x \in R_{(2,1)}} f(x, 2, 1)$, where $R_{(2,1)} = \{x \in \mathbb{R} \mid (x, 2, 1) \in E_{\underline{Z}}\} = \{1, 2\}$. Hence, $P(Y = 2, Z = 1) = f(1, 2, 1) + f(2, 2, 1) = \frac{1}{12}$. Therefore,

$$f_{X|(Y,Z)}(x|(2,1)) = \begin{cases} \frac{x}{3}, & \text{if } x \in \{1,2\} \\ 0, & \text{otherwise} \end{cases}$$

(2) The conditional p.m.f. of X, given $Y = 3$, is

$$f_{(X,Z)|Y}((x,z)|3) = \frac{f(x,3,z)}{P(Y=3)}$$
$$= \begin{cases} \frac{3xz}{72P(Y=3)}, \text{ if } x \in E_{X,Z)|Y=3} = \{(x,z) \in \mathbb{R} \mid (x,3,z) \in E_{\underline{Z}}\} \\ 0, \text{ otherwise} \end{cases}$$
$$= \begin{cases} \frac{3xz}{72P(Y=3)}, \text{ if } (x,z) \in \{1,2\} \times \{1,3\} \\ 0, \text{ otherwise} \end{cases}$$

Now, $P(Y=3) = \sum_{(x,z)\in R_3} f(x,3,z)$, where $R_3 = \{(x,z) \in \mathbb{R} \mid (x,3,z) \in E_{\underline{Z}}\} = \{1,2\} \times \{1,3\}$. Hence, $P(Y=3) = f(1,3,1)+f(1,3,3)+f(2,3,1)+f(2,3,3) = \frac{1}{2}$.

Therefore,

$$f_{(X,Z)|Y}((x,z)|3) = \begin{cases} \frac{xz}{12}, \text{ if } x \in \{1,2\} \times \{1,3\} \\ 0, \text{ otherwise} \end{cases}$$

2. Independent random variables

Definition 5. *The random variables $X_1, X_2, \ldots, X_n$ are said to be independent if for any sub-collection $\{X_{i_1}, X_{i_2}, \ldots, X_{i_k}\}$, $2 \le k \le n$, we have*

$$F_{X_{i_1},\ldots,X_{i_k}}(x_1, x_2, \cdots, x_k) = \prod_{j=1}^{k} F_{X_{i_j}}(x_j), \ \forall \ (x_1, x_2, \cdots, x_k) \in \mathbb{R}^k$$

where $F_{X_{i_1},\ldots,X_{i_k}}$ is the joint c.d.f. of $(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ and $F_{X_{i_j}}$ is the marginal c.d.f. of X_{i_j}, for $1 \le j \le k$.

Theorem 6. *Let $\underline{X} = (X_1, X_2, \ldots, X_n) : \mathcal{S} \longrightarrow \mathbb{R}^n$ be a $n-$dimensional ($n \ge 2$) random vector with joint c.d.f. $F_{\underline{X}}$. Let F_{X_i} be the marginal c.d.f. of X_i, for $1 \le i \le n$. Then the random variables $X_1, X_2, \ldots, X_n$ are independent if and only if*

$$F_{\underline{X}}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \ \forall \ (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n.$$

Theorem 7. *Let $\underline{X} = (X_1, X_2, \ldots, X_n) : \mathcal{S} \longrightarrow \mathbb{R}^n$ be a $n-$dimensional ($n \ge 2$) random vector of either discrete or continuous type. Let $f_{\underline{X}}$ be the joint p.m.f. (or p.d.f.) of $\underline{X}$ and f_{X_i} be the marginal p.m.f. (or p.d.f.) of random variable X_i, for $1 \le i \le n$. Then*

(1) *the random variables $X_1, X_2, \ldots, X_n$ are independent if and only if*

$$f_{\underline{X}}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \ \forall \ (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n.$$

(2) *the random variables $X_1, X_2, \ldots, X_n$ are independent $\Rightarrow E_{\underline{X}} = \prod_{i=1}^{n} E_{X_i}$, where $E_{\underline{X}}$ is the support of random vector $\underline{X}$ and E_{X_i} is the support of random variable X_i, for $1 \le i \le n$.*

Theorem 8. *Let $X_1, X_2, \ldots, X_n$ be the independent random variables.*

(1) *Let $\psi_i : \mathbb{R} \longrightarrow \mathbb{R}$ be a function such that $\psi_i(A) \in \mathbb{B}_{\mathbb{R}}$, for all $A \in \mathbb{B}_{\mathbb{R}}$, for $i = 1, 2, \cdots, n$. Then the random variables $\psi_1(X_1), \psi_2(X_2), \ldots, \psi_n(X_n)$ are independent.*

(2) *For $A_i \in \mathbb{B}_{\mathbb{R}}$, $i = 1, 2, \cdots, n$, we have*

$$P(\{X_i \in A_i, i = 1, 2, \cdots, n\}) = \prod_{i=1}^{n} P(\{X_i \in A_i).$$

Remark 9. *$\underline{X} = (X_1, X_2)$ be a random vector of either discrete or continuous type. Let $D = \{x_2 \in \mathbb{R} \mid f_{X_1|X_2}(.|x_2)$ is defined$\}$. Then for $x_2 \in D$, X_1 and X_2 are independent if and only if $f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1)$, for all $x_1 \in \mathbb{R}$,*

i.e,

X_1 and X_2 are independent if and only if $\forall\ x_2 \in D$, the conditional distribution of X_1, given $X_2 = x_2$, is the same as unconditional distribution of X_1.

Example 10. *Let $\underline{Z} = (X, Y, Z)$ be a random vector with joint p.m.f.*

$$f(x, y, z) = \begin{cases} \frac{xyz}{72}, & \text{if } (x, y, z) \in \{1, 2\} \times \{1, 2, 3\} \times \{1, 3\} \\ 0, & \text{otherwise} \end{cases}$$

(1) *Are X, Y and Z independent random variables?*
(2) *Are X and Z independent random variables?*

Solution:

(1) The supports of X, Y and Z are

$$E_X = \{x \in \mathbb{R} \mid (x, y, z) \in E_{\underline{Z}} \text{ for some } (y, z) \in \mathbb{R}^2\} = \{1, 2\}$$
$$E_Y = \{y \in \mathbb{R} \mid (x, y, z) \in E_{\underline{Z}} \text{ for some } (x, z) \in \mathbb{R}^2\} = \{1, 2, 3\}$$

and

$$E_Z = \{z \in \mathbb{R} \mid (x, y, z) \in E_{\underline{Z}} \text{ for some } (x, y) \in \mathbb{R}^2\} = \{1, 3\},$$

respectively. For $x \in E_X$, $R_x = \{(y, z) \in \mathbb{R}^2 \mid (x, y, z) \in E_{\underline{Z}}\} = \{1, 2, 3\} \times \{1, 3\}$. So the marginal p.m.f. of X is

$$f_X(x) = \begin{cases} \sum\limits_{(y,z) \in R_x} f(x, y, z), & \text{if } x \in E_X \\ 0, & \text{otherwise} \end{cases}$$
$$= \begin{cases} \frac{x}{3}, & \text{if } x \in \{1, 2\} \\ 0, & \text{otherwise} \end{cases}$$

Similarly the marginal p.m.f. of Y and Z are

$$f_Y(y) = \begin{cases} \frac{y}{6}, & \text{if } y \in \{1, 2, 3\} \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_Z(z) = \begin{cases} \frac{z}{4}, & \text{if } y \in \{1, 3\} \\ 0, & \text{otherwise} \end{cases}$$

respectively. Clearly $f(x, y, z) = f_X(x) f_Y(y) f_Z(z)$, for all $(x, y, z) \in \mathbb{R}^3$. Thus X, Y and Z are independent.

(2) Let $\underline{X} = (X, Y)$. The support of $\underline{X}$ is $E_{\underline{X}} = \{(x, z) \in \mathbb{R}^2 \mid (x, y, z) \in E_{\underline{Z}}$ for some $y \in \mathbb{R}\} = \{1, 2\} \times \{1, 3\}$. For $(x, z) \in E_{\underline{X}}$, $R_{(x,z)} = \{x \in \mathbb{R} \mid (x, y, z) \in E_{\underline{Z}}\} = \{1, 2, 3\}$.

So the marginal p.m.f. of $\underline{X}$ is

$$f_{\underline{X}}(x,z) = \begin{cases} \sum\limits_{y \in R_{(x,z)}} f(x,y,z), \text{ if } (x,z) \in E_{\underline{X}} \\ 0, \text{ otherwise} \end{cases}$$
$$= \begin{cases} \frac{xz}{12}, \text{ if } (x,z) \in \{1,2\} \times \{1,3\} \\ 0, \text{ otherwise} \end{cases}$$

Thus $f_{\underline{X}}(x,z) = f_X(x) f_Z(z)$, for all $(x,z) \in \mathbb{R}^2$. Thus X and Z are independent.

Example 11. *Let $\underline{Z} = (X,Y)$ be a random vector with joint p.d.f.*

$$f_{\underline{Z}}(x,y) = \begin{cases} \frac{1}{x}, \text{ if } 0 < y < x < 1 \\ 0, \text{ otherwise.} \end{cases}$$

Are X and Y independent?

Solution: By Example 7 of Lecture 14, the marginal p.d.f. of X and Y are

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

and

$$f_Y(y) = \begin{cases} -\ln y, \text{ if } 0 < y < 1 \\ 0, \text{ otherwise} \end{cases}$$

Clearly, $f_{\underline{Z}}(x,y) \neq f_X(x) f_Y(y)$. Hence, X and Y are not independent.

Alternative solution: The support of $\underline{Z}$ is $E_{\underline{Z}} = \{(x,y) \in \mathbb{R}^2 \mid 0 < y < x < 1\}$, and the support of X and Y are $(0,1)$. Hence, $E_{\underline{Z}} \neq E_X \times E_Y$. Therefore, X and Y are not independent.

Moments, Covariance and Correlation Coefficient

Let $\underline{X} = (X_1, X_2, \ldots, X_n)$ be a $n-$dimensional ($n \geq 2$) random vector and $\psi : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a function such that $\psi^{-1}(A) \in \mathbb{B}_{\mathbb{R}^n}$, for all $A \in \mathbb{B}_{\mathbb{R}}$. Suppose $E(\psi(\underline{X}))$ is finite.

(1) If $\underline{X}$ is of discrete type with joint p.m.f. $f_{\underline{X}}$ and support $E_{\underline{X}}$, then

$$E(\psi(\underline{X})) = \sum_{(x_1,x_2,\ldots,x_n)\in E_{\underline{X}}} \psi(x_1, x_2, \ldots, x_n)) f_{\underline{X}}(x_1, x_2, \ldots, x_n).$$

(2) If $\underline{X}$ is of continuous type with joint p.d.f. $f_{\underline{X}}$, then

$$E(\psi(\underline{X})) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \psi(x_1, x_2, \ldots, x_n) f_{\underline{X}}(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n.$$

(3) For nonnegative integers $k_1, k_2, \ldots, k_n$, let $\psi(x_1, x_2, \ldots, x_n) = x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$. Then

$$\mu'_{k_1,k_2,\ldots,k_n} = E(\psi(\underline{X})) = E(X_1^{k_1} X_2^{k_2} \cdots X_n^{k_n}),$$

provided it is finite, is called the joint moment of order $k_1 + k_2 + \cdots + k_n$ of $\underline{X} = (X_1, X_2, \ldots, X_n)$.

(4) For $n = 2$, let $\psi(x_1, x_2) = (x_1 - E(X_1))(x_2 - E(X_2))$. Then

$$Cov(X_1, X_2) = E\Big((X_1 - E(X_1))(X_2 - E(X_2))\Big),$$

provided it is finite, is called the covariance between X_1 and X_2.

Note: By the definition of covariance, it is easy to see

$$Cov(X_1, X_1) = Var(X_1);$$
$$Cov(X_1, X_2) = Cov(X_2, X_1);$$
$$Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2).$$

Theorem 1. *Let $\underline{X} = (X_1, X_2)$ and $\underline{Y} = (Y_1, Y_2)$ be two random vectors and a_1, a_2, b_1, b_2 be real constants. Then, provided the involved expectations are finite,*

(1) $E(a_1X_1 + a_2X_2) = a_1E(X_1) + a_2E(X_2)$;

(2) $Cov(a_1X_1+a_2X_2, b_1Y_1+b_2Y_2) = a_1b_1\ Cov(X_1, Y_1)+a_1b_2\ Cov(X_1, Y_2)+a_2b_1\ Cov(X_2, Y_1)+ a_2b_2\ Cov(X_2, Y_2) = \sum_{i=1}^{2}\sum_{j=1}^{2} a_ib_jCov(X_i, Y_j)$.

In particular,
$Var(a_1X_1 + a_2X_2) = Cov(a_1X_1 + a_2X_2, a_1X_1 + a_2X_2) = a_1^2Var(X_1) + a_2^2Var(X_2) + 2a_1a_2Cov(X_1, X_2)$.

Proof. (1) Suppose $\underline{X}$ is continuous type with joint p.d.f. $f_{\underline{X}}$. Let $\psi(x_1, x_2) = a_1x_1 + a_2x_2$. Then

$$\begin{aligned} E(a_1X_1 + a_2X_2) &= E(\psi(\underline{X})) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (a_1x_1 + a_2x_2) f_{\underline{X}}(x_1, x_2)\, dx_1\, dx_2 \\ &= a_1 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f_{\underline{X}}(x_1, x_2)\, dx_1\, dx_2 + a_2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f_{\underline{X}}(x_1, x_2)\, dx_1\, dx_2 \end{aligned}$$

By taking $\psi_1(x_1, x_2) = x_1$ and $\psi_2(x_1, x_2) = x_2$, we have

$$E(X_1) = E(\psi_1(\underline{X})) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f_{\underline{X}}(x_1, x_2)\, dx_1\, dx_2$$

and

$$E(X_2) = E(\psi_2(\underline{X})) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f_{\underline{X}}(x_1, x_2)\, dx_1\, dx_2.$$

Thus,

$$E(a_1X_1 + a_2X_2) = a_1E(X_1) + a_2E(X_2).$$

Similarly, we can prove for discrete type random vector.

(2)

$$\begin{aligned}
&Cov(a_1X_1 + a_2X_2, b_1Y_1 + b_2Y_2) \\
&\qquad = Cov(\sum_{i=1}^{2} a_iX_i, \sum_{j=1}^{2} b_jY_j) \\
&\qquad = E\Big(\big(\sum_{i=1}^{2} a_iX_i - E(\sum_{i=1}^{2} a_iX_i)\big)\big(\sum_{j=1}^{2} b_jY_j - E(\sum_{j=1}^{2} b_jY_j)\big)\Big) \\
&\qquad = E\Big(\big(\sum_{i=1}^{2} a_i(X_i - E(X_i))\big)\big(\sum_{j=1}^{2} b_j(Y_j - E(Y_j))\big)\Big) \text{ (by (1))} \\
&\qquad = E\Big(\sum_{i=1}^{2}\sum_{j=1}^{2} a_ib_j\big(X_i - E(X_i)\big)\big(Y_j - E(Y_j)\big)\Big) \\
&\qquad = \sum_{i=1}^{2}\sum_{j=1}^{2} a_ib_jE\Big(\big(X_i - E(X_i)\big)\big(Y_j - E(Y_j)\big)\Big) \\
&\qquad = \sum_{i=1}^{2}\sum_{j=1}^{2} a_ib_jCov(X_i, Y_j).
\end{aligned}$$

□

Remark 2. *In general, we have*

(1) $E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n)$;
(2) $Cov(\sum_{i=1}^{n_1} a_iX_i, \sum_{j=1}^{n_2} b_jY_j) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} a_ib_jCov(X_i, Y_j)$.
In particular,
$Var(\sum_{i=1}^{n_1} a_iX_i) = \sum_{i=1}^{n_1} a_iVar(X_i) + 2\sum\sum_{1\le i<j\le n_1} Cov(X_i, X_j)$

Theorem 3. *Let $X_1, X_2, \ldots, X_n$ be the independent random variables. Let $\psi_i : \mathbb{R} \longrightarrow \mathbb{R}$ be a function such that $\psi_i^{-1}(A) \in \mathbb{B}_{\mathbb{R}}$, for all $A \in \mathbb{B}_{\mathbb{R}}$, for $i = 1, 2, \cdots, n$. Then*

$$E\Big(\prod_{i=1}^{n} \psi_i(X_i)\Big) = \prod_{i=1}^{n} E\Big(\psi_i(X_i)\Big),$$

provided the involved expectations are finite.

Proof. We will prove the theorem for $n = 2$ and continuous random vector. Suppose $\underline{X} = (X_1, X_2)$ is a continuous type random vector with joint p.d.f. $f_{\underline{X}}$. Consider the function

$\psi(x_1, x_2) = \psi_1(x_1)\psi_2(x_2)$. Then

$$\begin{aligned}
E(\psi_1(X_1)\psi_2(X_2)) &= E(\psi(\underline{X})) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi_1(x_1)\psi_2(x_2) f_{\underline{X}}(x_1, x_2)\, dx_1\, dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi_1(x_1)\psi_2(x_2) f_{X_1}(x_1) f_{X_2}(x_2)\, dx_1\, dx_2 \text{ (since } X_1 \text{ and } X_2 \text{ are independent)} \\
&= \Big(\int_{-\infty}^{\infty} \psi_1(x_1) f_{X_1}(x_1)\, dx_1\Big)\Big(\int_{-\infty}^{\infty} \psi_2(x_2) f_{X_2}(x_2)\, dx_2\Big) \\
&= E(\psi_1(X_1))E(\psi_2(X_2))
\end{aligned}$$

□

Corollary 4. *Let $X_1, X_2, \ldots, X_n$ be the independent random variables. Then*

$$Cov(X_i, X_j) = 0,\ \forall\ i \neq j$$

and for real constants $a_1, a_2, \ldots, a_n$,

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i Var(X_i),$$

provided the involved expectations are finite.

Proof. Fix $i, j \in \{1, 2, \ldots, n\}, i \neq j$. Then by Theorem 3, we have

$$\begin{aligned}
&E(X_iX_j) = E(X_i)E(X_j) \\
\Rightarrow &Cov(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j) = 0
\end{aligned}$$

Since $Cov(X_i, X_j) = 0,\ \forall\ i \neq j$, by Remark 2,

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i Var(X_i).$$

□

Definition 5. (1) *The correlation coefficient between random variables X and Y is defined by*

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

provided $0 < Var(X), Var(Y) < \infty$.

(2) *The random variables X and Y are said to be uncorrelated if $Cov(X, Y) = 0$.*

Note: By definition, it is clear that if X and Y are independent random variables, then they are uncorrelated but converse need not be true.

Theorem 6. *Let X and Y be two random variables. Then, provided the involved expectations are finite,*

(1) $(E(XY))^2 \leq E(X^2)E(Y^2)$. *Moreover, $(E(XY))^2 = E(X^2)E(Y^2)$ if and only if $P(Y = cX) = 1$ or $P(X = cY) = 1$, for some $c \in \mathbb{R}$.*
This inequality is know as Cauchy-Schwarz inequality for random variables.

(2) $|\rho(X, Y)| \leq 1$. *To prove it, apply (1) on random variables $X' = X - E(X)$ and $Y' = Y - E(Y)$.*

Example 7. *Let $\underline{Z} = (X, Y)$ be a random vector of discrete type with joint p.m.f.*

$$f(x,y) = \begin{cases} p_1, & \text{if } (x,y) = (-1,1) \\ p_2, & \text{if } (x,y) = (0,0) \\ p_1, & \text{if } (x,y) = (1,1) \\ 0, & \text{otherwise} \end{cases}$$

where $p_1, p_2 \in (0,1)$ and $2p_1 + p_2 = 1$.

Then the support of $\underline{Z}, X$ and Y are

$$E_{\underline{Z}} = \{(-1,1),(0,0),(1,1)\}$$
$$E_X = \{-1,0,1\}$$
and
$$E_Y = \{0,1\},$$

respectively. Clearly $E_{\underline{Z}} \neq E_X \times E_Y$. So, X and Y are not independent.

Now,

$$E(XY) = \sum_{(x,y)\in E_{\underline{Z}}} xyf(x,y) = 0;$$
$$E(X) = \sum_{(x,y)\in E_{\underline{Z}}} xf(x,y) = 0;$$
$$E(Y) = \sum_{(x,y)\in E_{\underline{Z}}} yf(x,y) = 2p_1;$$
$$\Rightarrow Cov(X,Y) = E(XY) - E(X)E(Y) = 0 \Rightarrow \rho(X,Y) = 0$$

This shows that X and Y are uncorrelated but not independent.

We can also show that X and Y are not independent by another way.

The marginal p.m.f. of X is

$$f_X(x) = \begin{cases} \sum\limits_{y\in R_x} f(x,y), & \text{if } x \in \{-1,0,1\} \\ 0, & \text{otherwise} \end{cases}$$
$$= \begin{cases} p_1, & \text{if } x = -1 \\ p_2, & \text{if } x = 0 \\ p_1, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

Similarly, the marginal p.m.f. of Y is

$$f_Y(y) = \begin{cases} \sum\limits_{x\in R_y} f(x,y), & \text{if } y \in \{0,1\} \\ 0, & \text{otherwise} \end{cases}$$
$$= \begin{cases} p_2, & \text{if } x = 0 \\ 2p_1, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

Since $f(-1,1) \neq f_X(-1)f_Y(1)$, X and Y are not independent.

Example 8. *Let $\underline{Z} = (X, Y)$ be a random vector of continuous type with joint p.d.f.*

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

Now,

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dx\,dy = \int_{0}^{1}\int_{-x}^{x} xy\,dy\,dx = 0;$$

$$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\,dx\,dy = \int_{0}^{1}\int_{-x}^{x} x\,dy\,dx = \frac{2}{3};$$

$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\,dx\,dy = \int_{0}^{1}\int_{-x}^{x} y\,dy\,dx = 0;$$

$$\Rightarrow Cov(X,Y) = E(XY) - E(X)E(Y) = 0 \Rightarrow \rho(X,Y) = 0$$

Thus X and Y are uncorrelated.

The marginal p.d.f. of X is

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

$$= \begin{cases} \int_{-x}^{x} dy, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} 2x, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Similarly, the marginal p.d.f. of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

$$= \begin{cases} \int_{|y|}^{1} dx, & \text{if } -1 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} 1 - |y|, & \text{if } -1 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Since $f(x,y) \neq f_X(x)f_Y(y)$, X and Y are not independent.

We can also show that X and Y are not independent by another way. Then the support of $\underline{Z}, X$ and Y are

$$E_{\underline{Z}} = \{(x,y) \in \mathbb{R}^2 \mid 0 < |y| \leq x < 1\}$$

$$E_X = (0,1)$$

and

$$E_Y = (-1,1),$$

respectively. Clearly $E_{\underline{Z}} \neq E_X \times E_Y$. So, X and Y are not independent.

This example also shows that X and Y are uncorrelated but not independent.

Conditional Expectation and Variance

Definition 1. *Let (X,Y) be a random vector and $h:\mathbb{R}\longrightarrow\mathbb{R}$ be a function such that $h^{-1}(A)\in\mathbb{B}_{\mathbb{R}}$, for all $A\in\mathbb{B}_{\mathbb{R}}$. Then*

(1) *the conditional expectation of $h(X)$, given Y, written as $E[h(X)|Y]$, is an random variable that takes the value $E[h(X)|Y=y]$, defined by*

$$E[h(X)|Y=y]=\begin{cases}\sum\limits_{x\in E_{X|Y=y}} h(x)P(X=x|Y=y), & \text{if } (X,Y) \text{ is of discrete type and } P(Y=y)>0\\ \int\limits_{-\infty}^{\infty} h(x)f_{X|Y}(x|y)\,dx, & \text{if } (X,Y) \text{ is of continuous type and } f_Y(y)>0\end{cases}$$

(2) *the conditional variance of $h(X)$, given Y, written as $Var[h(X)|Y]$, is an random variable that takes the value $Var[h(X)|Y=y]$, defined by*

$$\begin{aligned}Var[h(X)|Y=y]&=E[\big(h(X)-E[h(X)|Y=y]\big)^2|Y=y]\\&=E[\big(h(X)\big)^2|Y=y]-(E[h(X)|Y=y])^2\end{aligned}$$

Remark 2. (1) *For any constant c, $E[c|Y]=c$.*

(2) *Let $h_i:\mathbb{R}\longrightarrow\mathbb{R}$ be a function such that $h_i^{-1}(A)\in\mathbb{B}_{\mathbb{R}}$, for all $A\in\mathbb{B}_{\mathbb{R}}$, for $i=1,2$. Then*

$$E[a_1h_1(X)+a_2h_2(X)|Y]=a_1E[h_1(X)|Y]+a_2E[h_2(X)|Y],$$

for any constants a_1,a_2.

(3) *If X and Y are independent, then*

$$E[h(X)|Y]=E(h(X)) \text{ and } Var[h(X)|Y]=Var(h(X)).$$

(4) *If $P(X\geq 0)=1$, then $E[X|Y]\geq 0$.*

(5) *If $P(X_1\geq X_2)=1$, then $E[X_1|Y]\geq E[X_1|Y]$.*

Theorem 3. (1) *Let $E(h(X))$ exist. Then*

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

(2) ***The conditional Variance Formula:***

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

Proof. Let (X,Y) be of the discrete type. Then

$$\begin{aligned}E(E[h(X)|Y])&=\sum_y E[h(X)|Y=y]P(Y=y)\\&=\sum_y\left[\sum_x h(x)P(X=x|Y=y)\right]P(Y=y)\\&=\sum_y\left[\sum_x h(x)P(X=x,Y=y)\right]\\&=\sum_x\left[\sum_y h(x)P(X=x,Y=y)\right]\\&=\sum_x h(x)P(X=x)\\&=E(h(X)).\end{aligned}$$

(2)

$$\begin{aligned} E(Var[h(X)|Y]) &= E(E[\big(h(X)\big)^2|Y] - (E[h(X)|Y])^2) \\ &= E(E[\big(h(X)\big)^2|Y]) - E((E[h(X)|Y])^2)) \\ &= E(\big(h(X)\big)^2) - E((E[h(X)|Y])^2)) \end{aligned}$$

and

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

Thus $Var(E[h(X)|Y]) + E(Var[h(X)|Y]) = E(\big(h(X)\big)^2) - (E(h(X)))^2 = Var(h(X))$.

□

Example 4. *Let $\underline{Z} = (X, Y, Z)$ be a random vector with joint p.m.f.*

$$f(x,y,z) = \begin{cases} \frac{xyz}{72}, & \text{if } (x,y,z) \in \{1,2\} \times \{1,2,3\} \times \{1,3\} \\ 0, & \text{otherwise} \end{cases}$$

(1) *Let $Y_1 = 2X - Y + 3Z$ and $Y_2 = X - 2Y + Z$. Find the correlation coefficient between Y_1 and Y_2.*
(2) *For a fixed $y \in \{1,2,3\}$, find $E[Y_3|Y = y]$ and $Var[Y_3|Y = y]$, where $Y_3 = XZ$.*

Solution:

(1) By Example 10 of Lecture 16, we know that the marginal p.m.f. of X, Y and Z are

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

$$f_Y(y) = \begin{cases} \frac{y}{6}, & \text{if } y \in \{1,2,3\} \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_Z(z) = \begin{cases} \frac{z}{4}, & \text{if } y \in \{1,3\} \\ 0, & \text{otherwise} \end{cases}$$

respectively. Also X, Y, Z are independent. Therefore, $Cov(X,Y) = Cov(X,Z) = Cov(Y,Z) = 0$. Hence,

$$Cov(Y_1, Y_2) = 2Var(X) + 2Var(Y) + 3Var(Z);$$

$$Var(Y_1) = 4Var(X) + Var(Y) + 9Var(Z);$$

and

$$Var(Y_2) = Var(X) + 4Var(Y) + Var(Z).$$

By a simple calculation, we have

$$E(X) = \frac{5}{3}, E(Y) = \frac{7}{3} \text{ and } E(Z) = \frac{5}{2};$$

$$E(X^2) = 3, E(Y^2) = 6 \text{ and } E(Z^2) = 7;$$

$$Var(X) = \frac{2}{9}, Var(Y) = \frac{5}{9} \text{ and } Var(Z) = \frac{3}{4}$$

Therefore,

$$Cov(Y_1, Y_2) = \frac{137}{36}, Var(Y_1) = \frac{295}{36} \text{ and } Var(Y_2) = \frac{115}{36}.$$

Thus,

$$\rho(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1)Var(Y_2)}} = \frac{137}{\sqrt{295}\sqrt{115}}.$$

(2) As we know that X, Y, Z are independent, it follows that (X, Z) and Y are independent. Thus, $Y_3 = XZ$ and Y are independent. Therefore, $E[Y_3|Y = y] = E(Y_3) = E(X)E(Z) = \frac{25}{6}$ and

$$\begin{aligned} Var[Y_3|Y = y] &= Var(Y_3) \\ &= Var(E[XZ|Z]) + E(Var[XZ|Z]) \\ &= Var(ZE[X|Z]) + E(Z^2 Var[X|Z]) \\ &= Var(ZE(X)) + E(Z^2 Var(X)) \\ &= Var(\frac{5}{3}Z) + E(\frac{2}{9}Z^2) \\ &= \frac{25}{9}Var(Z) + \frac{2}{9}E(Z^2) \\ &= \frac{131}{36}. \end{aligned}$$

Example 5. *Let $\underline{Z} = (X, Y)$ be a random vector with joint p.d.f.*

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

For a fixed $0 < x < 1$, find $E[Y|X = x]$ and $Var[Y|X = x]$, and for a fixed $0 < y < 1$, find $E[X|Y = y]$ and $Var[X|Y = y]$.

Solution: The marginal p.d.f. of X and Y are

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_x^1 2\, dy = \begin{cases} 2(1-x), & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^y 2\, dx = \begin{cases} 2y, & \text{if } 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

respectively. Hence, the conditional p.d.f. of Y, given $X = x$ and the conditional p.d.f. of X, given $Y = y$ are

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \begin{cases} \frac{1}{1-x}, & \text{if } x < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \begin{cases} \frac{1}{y}, & \text{if } 0 < x < y \\ 0, & \text{otherwise} \end{cases}$$

Thus,

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy = \int_x^1 \frac{y}{1-x}\, dy = \frac{1+x}{2};$$

$$E[Y^2|X=x] = \int_{-\infty}^{\infty} y^2 f_{Y|X}(y|x)\,dy = \int_{x}^{1} \frac{y^2}{1-x}\,dy = \frac{1+x+x^2}{3};$$

$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,dx = \int_{0}^{y} \frac{x}{y}\,dy = \frac{y}{2};$$

$$E[X^2|Y=y] = \int_{-\infty}^{\infty} x^2 f_{X|Y}(x|y)\,dx = \int_{0}^{y} \frac{x^2}{y}\,dy = \frac{y^2}{3}.$$

Hence,

$$Var[Y|X=x] = E[Y^2|X=x] - (E[Y|X=x])^2 = \frac{1+x+x^2}{3} - \frac{1+2x+x^2}{4} = \frac{x^2-2x+1}{12};$$

and

$$Var[X|Y=y] = E[X^2|Y=y] - (E[X|Y=y])^2 = \frac{y^2}{3} - \frac{y^2}{4} = \frac{y^2}{12}.$$

Example 6. *Suppose that the expected number of accidents per week at an industrial plant is four. Suppose also that the numbers of workers injured in each accident are independent random variables with a common mean of 2. Assume also that the number of workers injured in each accident is independent of the number of accidents that occur. What is the expected number of injuries during a week?*

Solution: Let N denote the number of accidents and X_i the number of workers injured in the $i-$th accident, $i = 1, 2, \ldots$, then the total number of injuries can be expressed as $\sum_{i=1}^{N} X_i$. Now, $E(\sum_{i=1}^{N} X_i) = E(E[\sum_{i=1}^{N} X_i|N])$.

But $E[\sum_{i=1}^{N} X_i|N=n] = E[\sum_{i=1}^{n} X_i|N=n] = E(\sum_{i=1}^{n} X_i) = nE(X_i)$ (since X_i and N are independent, and X_i has common mean). Thus, $E[\sum_{i=1}^{N} X_i|N] = NE(X_i)$. Therefore, $E(\sum_{i=1}^{N} X_i) = E(NE(X_i)) = E(N)E(X_i) = 8$.

Joint Moment generating function

Let $\underline{X} = (X_1, X_2, \cdots, X_n)$ be a $n-$ dimensional random vector and let $A = \{(t_1, t_2, \cdots, t_n) \in \mathbb{R}^n \mid E(e^{\sum_{i=1}^{n} t_i X_i}) \text{ is finite}\}$. The function $M_{\underline{X}} : A \longrightarrow \mathbb{R}$, defined by

$$M_{\underline{X}}(\underline{t}) = E(e^{\sum_{i=1}^{n} t_i X_i}),\ \forall\ \underline{t} = (t_1, t_2, \cdots, t_n) \in A$$

is known as the joint moment generating function (j.m.g.f.) of the random vector $\underline{X}$ if $E(e^{\sum_{i=1}^{n} t_i X_i})$ is finite on a rectangle $(-\underline{a}, \underline{a}) \subseteq A$ for some $(a_1, a_2, \cdots, a_n) \in \mathbb{R}^n$, where $a_i > 0,\ i = 1, 2, \cdots, n$.

Note:

(1) $M_{\underline{X}}(\underline{0}) = 1$, where $\underline{0} = (0, 0, \cdots, 0)$.

(2) If $X_1, X_2, \cdots, X_n$ are independent, then $M_{\underline{X}}(\underline{t}) = E(e^{\sum_{i=1}^{n} t_i X_i}) = E(\prod_{i=1}^{n} e^{t_i X_i}) = \prod_{i=1}^{n} E(e^{t_i X_i})$
$= \prod_{i=1}^{n} M_{X_i}(t_i),\ \forall\ \underline{t} = (t_1, t_2, \cdots, t_n) \in A$, where M_{X_i} is the m.g.f. of $X_i,\ i = 1, 2, \ldots, n$.

Theorem 1. *Let $\underline{X} = (X_1, X_2, \cdots, X_n)$ be a $n-$ dimensional random vector with the joint moment generating function (j.m.g.f.) $M_{\underline{X}}$ that is finite on a rectangle interval $(-\underline{a}, \underline{a}) = (-a_1, a_1) \times (-a_2, a_2) \times \cdots \times (-a_n, a_n) \subseteq \mathbb{R}^n$, where $a_i > 0,\ i = 1, 2, \ldots, n$. Then $M_{\underline{X}}$ possesses partial derivatives of all orders in $(-\underline{a}, \underline{a})$. Furthermore, for positive integers $k_1, k_2, \ldots, k_n$,*

$$E(X_1^{k_1} X_2^{k_2} \cdots X_n^{k_n}) = \left[\frac{\partial^{k_1+k_2+\cdots+k_n}}{\partial t_1^{k_1} \partial t_2^{k_2} \cdots \partial t_n^{k_n}} M_{\underline{X}}(\underline{t})\right]_{\underline{t}=\underline{0}},\ \text{where } \underline{t} = (t_1, t_2, \cdots, t_n) \text{ and } \underline{0} = (0, 0, \cdots, 0).$$

In particular,

$$E(X_i) = \left[\frac{\partial}{\partial t_i} M_{\underline{X}}(\underline{t})\right]_{\underline{t}=\underline{0}},\ i = 1, 2, \ldots, n;$$

$$E(X_i^m) = \left[\frac{\partial^m}{\partial t_i^m} M_{\underline{X}}(\underline{t})\right]_{\underline{t}=\underline{0}},\ i = 1, 2, \ldots, n;$$

$$Var(X_i) = \left[\frac{\partial^2}{\partial t_i^2} M_{\underline{X}}(\underline{t})\right]_{\underline{t}=\underline{0}} - \left(\left[\frac{\partial}{\partial t_i} M_{\underline{X}}(\underline{t})\right]_{\underline{t}=\underline{0}}\right)^2,\ i = 1, 2, \ldots, n;$$

and, for $i, j \in \{1, 2, \ldots, n\}, i \neq j$

$$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) = \left[\frac{\partial^2}{\partial t_i \partial t_j} M_{\underline{X}}(\underline{t})\right]_{\underline{t}=\underline{0}} - \left[\frac{\partial}{\partial t_i} M_{\underline{X}}(\underline{t})\right]_{\underline{t}=\underline{0}} \left[\frac{\partial}{\partial t_j} M_{\underline{X}}(\underline{t})\right]_{\underline{t}=\underline{0}}.$$

Also

$M_{\underline{X}}(0, \cdots, 0, t_i, 0, \cdots, 0) = E(e^{t_i X_i}) = M_{X_i}(t_i)$;

$M_{\underline{X}}(0, \cdots, 0, t_i, 0, \cdots, 0, t_j, 0, \cdots, 0) = E(e^{t_i X_i + t_j X_j}) = M_{X_i, X_j}(t_i, t_j),\ i, j \in \{1, 2, \ldots, n\}$,

provided the involved expectations are finite.

Definition 2. *Let $\underline{X}$ and $\underline{Y}$ be two $n-$ dimensional random vectors with joint c.d.f. $F_{\underline{X}}$ and $F_{\underline{Y}}$ respectively. We say that $\underline{X}$ and $\underline{Y}$ have the same distribution (or are identically distributed) if $F_{\underline{X}}(\underline{x}) = F_{\underline{Y}}(\underline{x}),\ \forall\ \underline{x} \in \mathbb{R}^n$. In this case,it is written as $\underline{X} \overset{\mathrm{d}}{=} \underline{Y}$.*

Theorem 3. (1) *Let $\underline{X}$ and $\underline{Y}$ be two $n-$ dimensional random vectors with joint p.m.f.'s $f_{\underline{X}}$ and $f_{\underline{Y}}$, respectively. Then, $\underline{X} \overset{\mathrm{d}}{=} \underline{Y}$ if and only if $f_{\underline{X}}(\underline{x}) = f_{\underline{Y}}(\underline{x}).\ \forall\ \underline{x} \in \mathbb{R}^n$.*

(2) *Let $\underline{X}$ and $\underline{Y}$ be two $n-$ dimensional continuous type random vectors. Then, $X \stackrel{\mathrm{d}}{=} Y$ if and only if there exist versions of joint p.d.f.'s $f_{\underline{X}}$ and $f_{\underline{Y}}$ of $\underline{X}$ and $\underline{Y}$, respectively, such that $f_{\underline{X}}(\underline{x}) = f_{\underline{Y}}(\underline{x})$. $\forall\ \underline{x} \in \mathbb{R}^n$.*

Theorem 4. *Let $\underline{X}$ and $\underline{Y}$ be two $n-$ dimensional random vectors of either discrete type or of continuous type with $\underline{X} \stackrel{\mathrm{d}}{=} \underline{Y}$. Then, for any function $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ with $h^{-1}(A) \in \mathbb{B}_{\mathbb{R}^n}$, for every $A \in \mathbb{B}_{\mathbb{R}}$, we have*

$$h(\underline{X}) \stackrel{\mathrm{d}}{=} h(\underline{Y})$$

and

$$E(h(\underline{X})) = E(h(\underline{Y})),$$

provided the expectations are finite.

Theorem 5. *X_1 and X_2 are independent random variables if and only if $M_{X_1,X_2}(t_1, t_2) = M_{X_1,X_2}(t_1, 0,)M_{X_1,X_2}(0, t_2)$, for all $(t_1, t_2) \in \mathbb{R}^2$.*

Theorem 6. *Let $\underline{X}$ and $\underline{Y}$ be two $n-$ dimensional random vectors of either discrete type or of continuous type with having joint m.g.f.'s $M_{\underline{X}}$ and $M_{\underline{Y}}$, respectively that are finite on a rectangle $(-\underline{a}, \underline{a}) \subseteq A$ for some $(a_1, a_2, \cdots, a_n) \in \mathbb{R}^n$, where $a_i > 0,\ i = 1, 2, \cdots, n$. Suppose that $M_{\underline{X}}(\underline{t}) = M_{\underline{Y}}(\underline{t})$, $\forall\ \underline{t} \in (-\underline{a}, \underline{a})$. Then $\underline{X} \stackrel{\mathrm{d}}{=} \underline{Y}$.*

Example 7. *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \sim Bin(n_i, \theta),\ 0 < \theta < 1, n_i\{1, 2, \cdots\}, i = 1, 2, \ldots, n$. Then show that*

$$\sum_{i=1}^{n} X_i \sim Bin(\sum_{i=1}^{n} n_i, \theta).$$

Solution: Let $Y = \sum_{i=1}^{n} X_i$. Then

$$\begin{aligned}
M_Y(t) &= E(e^{t\sum_{i=1}^{n} X_i}) \\
&= E(\prod_{i=1}^{n} e^{tX_i}) \\
&= \prod_{i=1}^{n} E(e^{tX_i}) \\
&= \prod_{i=1}^{n} M_{X_i}(t) \\
&= \prod_{i=1}^{n} (1 - \theta + \theta e^t) \\
&= (1 - \theta + \theta e^t)^{\sum_{i=1}^{n} n_i},\ \forall\ t \in \mathbb{R}
\end{aligned}$$

Since m.g.f. of $\mathrm{Bin}(\sum_{i=1}^{n} n_i, \theta)$ is $(1 - \theta + \theta e^t)^{\sum_{i=1}^{n} n_i}$, by Theorem 6, $Y \sim \mathrm{Bin}(\sum_{i=1}^{n} n_i, \theta)$.

Functions of several random variables

Let $\underline{X} = (X_1, X_2, \cdots, X_n)$ be a n-dimensional random vector. Let $g : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be a function such that $g^{-1}(A) \in \mathbb{B}_{\mathbb{R}^n}$, for all $A \in \mathbb{B}_{\mathbb{R}^m}$. Then $Y = g(\underline{X})$ is an m-dimensional random vector.

Suppose joint c.d.f. of $\underline{X} = (X_1, X_2, \cdots, X_n)$ is $F_{\underline{X}}$ and $m = 1$, i.e., $Y = g(\underline{X})$ is a random variable. Let $y \in \mathbb{R}$. Then

$$P(Y \leq y) = P(g(X_1, X_2, \cdots, X_n) \leq y)$$
$$= \begin{cases} \sum\limits_{\{(x_1,x_2,\cdots,x_n): g(X_1,X_2,\cdots,X_n) \leq y\}} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n), \text{ if } \underline{X} \text{ is of discrete type} \\ \int\limits_{\{(x_1,x_2,\cdots,x_n): g(X_1,X_2,\cdots,X_n) \leq y\}} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n, \text{ if } \underline{X} \text{ is of continuous type} \end{cases}$$

where f is the joint p.d.f. of $\underline{X} = (X_1, X_2, \cdots, X_n)$ in case of continuous type.

Example 1. *Let X_1, X_2 be independent uniform distributions $U(0,1)$. Find the c.d.f. of $Y = X_1 + X_2$ and hence find the p.d.f. of Y.*

Solution: The joint p.d.f. of (X_1, X_2) is

$$f(x_1, x_2) = \begin{cases} 1, \text{ if } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0, \text{ otherwise} \end{cases}$$

Now, the c.d.f. of Y is

$$F_Y(y) = P(Y \leq y)$$
$$= P(X_1 + X_2 \leq y)$$
$$= \iint\limits_{\{(x_1,x_2):\ x_1+x_2 \leq y\}} f(x_1, x_2) dx_1 dx_2$$
$$= \begin{cases} 0, \text{ if } y < 0 \\ \int\limits_0^y \Big(\int\limits_0^{y-x_1} dx_2 \Big) dx_1, \text{ if } 0 \leq y < 1 \\ 1 - \int\limits_{y-1}^1 \Big(\int\limits_{y-x_2}^1 dx_1 \Big) dx_2, \text{ if } 1 \leq y < 2, \text{ (since } P(X_1 + X_2 \leq y) = 1 - P(X_1 + X_2 > y)) \\ 1, \text{ if } y \geq 2 \end{cases}$$
$$= \begin{cases} 0, \text{ if } y < 0 \\ \frac{y^2}{2}, \text{ if } 0 \leq y < 1 \\ \frac{4y - y^2 - 2}{2}, \text{ if } 1 \leq y < 2 \\ 1, \text{ if } y \geq 2 \end{cases}$$

Hence, p.d.f. of Y is

$$f_Y(y) = \begin{cases} y, \text{ if } 0 \leq y < 1 \\ 2 - y, \text{ if } 1 \leq y < 2 \\ 0, \text{ otherwise} \end{cases}$$

Example 2. *Let $\underline{X} = (X_1, X_2)$ be a continuous random vector with joint p.d.f. is*

$$f(x_1, x_2) = \begin{cases} e^{-x_1}, \text{ if } 0 < x_2 \leq x_1 < \infty \\ 0, \text{ otherwise} \end{cases}$$

. Find p.d.f. of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$.

Solution: The c.d.f. of Y_1 is

$$\begin{aligned}
F_{Y_1}(y) &= P(Y_1 \le y) \\
&= P(X_1 + X_2 \le y) \\
&= \iint\limits_{\{(x_1,x_2):\ x_1+x_2\le y\}} f(x_1,x_2)dx_1dx_2 \\
&= \begin{cases} 0, \text{ if } y<0 \\ \int\limits_0^{\frac{y}{2}} \big(\int\limits_{x_2}^{y-x_2} dx_1 \big) dx_2, \text{ if } y \ge 0 \end{cases} \\
&= \begin{cases} 0, \text{ if } y<0 \\ (1-e^{-\frac{y}{2}})^2, \text{ if } y \ge 0 \end{cases}
\end{aligned}$$

Hence, p.d.f. of Y_1 is

$$f_Y(y) = \begin{cases} 0, \text{ if } y<0 \\ (1-e^{-\frac{y}{2}})e^{-\frac{y}{2}}, \text{ if } y \ge 0 \end{cases}$$

Now, the c.d.f. of Y_2 is

$$\begin{aligned}
F_{Y_2}(y) &= P(Y_2 \le y) \\
&= P(X_1 - X_2 \le y) \\
&= \iint\limits_{\{(x_1,x_2):\ x_1-x_2\le y\}} f(x_1,x_2)dx_1dx_2 \\
&= \begin{cases} 0, \text{ if } y<0 \\ \int\limits_0^{\infty} \big(\int\limits_{x_2}^{y+x_2} dx_1 \big) dx_2, \text{ if } y \ge 0 \end{cases} \\
&= \begin{cases} 0, \text{ if } y<0 \\ 1-e^{-y}, \text{ if } y \ge 0 \end{cases}
\end{aligned}$$

Hence, p.d.f. of Y_2 is

$$f_Y(y) = \begin{cases} 0, \text{ if } y<0 \\ e^{-y}, \text{ if } y \ge 0 \end{cases}$$

Transformation of Variables Technique:

Theorem 3. *Let $\underline{X} = (X_1, X_2, \cdots, X_n)$ be a discrete type random vector with support $E_{\underline{X}}$ and the p.m.f. $f_{\underline{X}}$. Let $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a function such that $g_i^{-1}(A) \in \mathbb{B}_{\mathbb{R}^n}$, for all $A \in \mathbb{B}_{\mathbb{R}}$ and $Y_i = g_i(\underline{X})$, $i = 1, 2, \cdots, k$. Define, for $\underline{y} = (y_1, y_2, \cdots, y_k) \in \mathbb{R}^k$, $A_{\underline{y}} = \{\underline{x} = (x_1, x_2, \cdots, x_n) \in E_{\underline{X}} \mid g_1(\underline{x}) \le y_1, \ldots, g_k(\underline{x}) \le y_k\}$ and $B_{\underline{y}} = \{\underline{x} = (x_1, x_2, \cdots, x_n) \in E_{\underline{X}} \mid g_1(\underline{x}) = y_1, \ldots, g_k(\underline{x}) = y_k\}$. Then the random vector $\underline{Y} = (Y_1, Y_2, \cdots, Y_k)$ is of discrete type with joint c.d.f.*

$$F_{\underline{Y}}(\underline{y}) = \sum_{\underline{x} \in A_{\underline{y}}} f_{\underline{X}}(\underline{x}),\ \underline{y} \in \mathbb{R}^k$$

and the p.m.f.

$$f_{\underline{Y}}(\underline{y}) = \sum_{\underline{x} \in B_{\underline{y}}} f_{\underline{X}}(\underline{x}),\ \underline{y} \in \mathbb{R}^k.$$

Example 4. *Let X_1, X_2 be independent random variables with $X_1 \sim Bin(n_1, \theta)$ and $X_2 \sim Bin(n_2, \theta)$, where $n_1, n_2 \in \mathbb{N}$. Without using the m.g.f. of $Y = X_1 + X_2$, find the p.m.f. of Y.*

Solution: The joint p.m.f. of $\underline{X} = (X_1, X_2)$ is given by

$$\begin{aligned} f_{\underline{X}}(x_1, x_2) &= f_{X_1}(x_1) f_{X_2}(x_2) \\ &= \begin{cases} \prod_{i=1}^{2} \binom{n_i}{x_i} \theta^{x_i} (1-\theta)^{n_i - x_i}, & \text{if } (x_1, x_2) \in \prod_{i=1}^{2} \{0, 1, \cdots, n_i\} \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} \left(\prod_{i=1}^{2} \binom{n_i}{x_i} \right) \theta^{x_1 + x_2} (1-\theta)^{(n_1+n_2)-(x_1+x_2)}, & \text{if } (x_1, x_2) \in \prod_{i=1}^{2} \{0, 1, \cdots, n_i\} \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

where $\prod_{i=1}^{2} \{0, 1, \cdots, n_i\} = \{0, 1, \cdots, n_1\} \times \{0, 1, \cdots, n_2\}$. Let $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be a function defined by $g(x_1, x_2) = x_1 + x_2$. Then $Y = g(\underline{X}) = X_1 + X_2$. Thus, for $y \notin \{0, 1, \cdots, n_1 + n_2\}$, $B_y = \{(x_1, x_2) \in E_{\underline{X}} \mid x_1 + x_2 = y\} = \emptyset$ and $f_Y(y) = P(Y = y) = 0$. Now, for $y \in \{0, 1, \cdots, n_1 + n_2\}$, we have

$$\begin{aligned} f_Y(y) = P(Y = y) &= \sum_{\underline{x} \in B_y} f_{\underline{X}}(x_1, x_2) \\ &= \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2}{}_{x_1+x_2=y} \left(\prod_{i=1}^{2} \binom{n_i}{x_i} \right) \theta^{x_1+x_2} (1-\theta)^{(n_1+n_2)-(x_1+x_2)} \\ &= \sum_{x_1=0}^{y} \binom{n_1}{x_1} \binom{n_2}{y - x_1} \theta^y (1-\theta)^{(n_1+n_2)-y} \\ &= \binom{n_1 + n_2}{y} \theta^y (1-\theta)^{(n_1+n_2)-y} \end{aligned}$$

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

$$f_Y(y) = \begin{cases} \binom{n}{y} \theta^y (1-\theta)^{n-y}, \text{if } y \in \{0, 1, \cdots, n\} \\ 0, \text{ otherwise} \end{cases}$$

where $n = n_1 + n_2$.

Theorem 5. *Let $\underline{X} = (X_1, X_2, \cdots, X_n)$ be an $n-$dimensional random vector of continuous type with joint p.d.f. $f_{\underline{X}}$.*

(1) *Let*

$$\begin{aligned} y_1 &= g_1(x_1, x_2, \ldots, x_n) \\ y_2 &= g_2(x_1, x_2, \ldots, x_n) \\ &\vdots \\ y_n &= g_n(x_1, x_2, \ldots, x_n) \end{aligned}$$

be a one-to-one mapping from $\mathbb{R}^n$ into $\mathbb{R}^n$ sending $(x_1, x_2, \ldots, x_n)$ to $(y_1, y_2, \ldots, y_n)$. That is, there exists the inverse transformation

$$\begin{aligned} x_1 &= h_1(y_1, y_2, \ldots, y_n) \\ x_2 &= h_2(y_1, y_2, \ldots, y_n) \\ &\vdots \\ x_n &= h_n(y_1, y_2, \ldots, y_n) \end{aligned}$$

defined over the range of the transformation.

In other words, the equations $y_i = g_i(x_1, x_2, \ldots, x_n)$, $1 \le i \le n$ have a unique solution $x_i = h_i(y_1, y_2, \ldots, y_n)$, $1 \le i \le n$.

(2) *Assume that both the mapping and its inverse are continuous.*

(3) *Assume that the partial derivatives*

$$\frac{\partial x_i}{\partial y_j}, 1 \le i \le n, 1 \le j \le n$$

exist and are continuous.

(4) *Assume that the Jacobian J of the inverse transformation*

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \neq 0$$

for $(y_1, y_2, \ldots, y_n)$ in the range of the transformation.

Then $\underline{Y} = (Y_1, Y_2, \cdots, Y_n) = (g_1(X_1, X_2, \cdots, X_n), g_2(X_1, X_2, \cdots, X_n), \cdots, g_n(X_1, X_2, \cdots, X_n))$ is a continuous random vector with joint p.d.f.

$$f_{\underline{Y}}(y_1, y_2, \cdots, y_n) = f_{\underline{X}}(h_1(y_1, y_2, \cdots, y_n), h_2(y_1, y_2, \cdots, y_n), \cdots, h_n(y_1, y_2, \cdots, y_n))|J|$$

Example 6. *Let X_1 and X_2 be independent uniform distributions $U(0,1)$. Find p.d.f. of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$.*

Solution: The joint p.d.f. of (X_1, X_2) is given by

$$f(x_1, x_2) = \begin{cases} 1, \text{ if } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0, \text{ otherwise} \end{cases}$$

Let $y_1 = x_1 + x_2$ and $y_2 = x_1 - x_2$. Then $x_1 = \frac{y_1+y_2}{2}$ and $x_2 = \frac{y_1-y_2}{2}$ and

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = -\frac{1}{2}$$

Therefore the joint p.d.f. of $\underline{Y} = (Y_1, Y_2)$ is

$$f_{\underline{Y}}(y_1, y_2) = f(\frac{y_1 + y_2}{2}, \frac{y_1 - y_2}{2})|J| = \begin{cases} \frac{1}{2}, \text{ if } 0 < y_1 + y_2 < 2, 0 < y_1 - y_2 < 2 \\ 0, \text{ otherwise} \end{cases}$$

Now, the marginal p.d.f. of Y_1 is

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} f_{\underline{Y}}(y_1, y_2) dy_2 \\ &= \begin{cases} \int_{-y_1}^{y_1} \frac{1}{2} dy_2, \text{ if } 0 < y \le 1 \\ \int_{y_1-2}^{2-y_1} \frac{1}{2} dy_2, \text{ if } 1 < y < 2 \\ 0, \text{ otherwise} \end{cases} \\ &= \begin{cases} y_1, \text{ if } 0 < y \le 1 \\ 2 - y_1, \text{ if } 1 < y < 2 \\ 0, \text{ otherwise} \end{cases} \end{aligned}$$

The marginal p.d.f. of Y_2 is

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{\underline{Y}}(y_1, y_2)dy_1$$

$$= \begin{cases} \int_{-y_2}^{2+y_2} \frac{1}{2}dy_2, \text{ if } -1 < y_2 \leq 0 \\ \int_{y_2}^{2-y_2} \frac{1}{2}dy_2, \text{ if } 0 < y_2 < 1 \\ 0, \text{ otherwise} \end{cases}$$

$$= \begin{cases} 1 + y_2, \text{ if } -1 < y_2 \leq 0 \\ 1 - y_2, \text{ if } 0 < y_2 < 1 \\ 0, \text{ otherwise} \end{cases}$$

Example 7. *Let X_1 and X_2 be independent exponential distributions Exp(λ). Show that $\frac{X_1}{X_1+X_2} \sim U(0,1)$.*

Solution: The joint p.d.f. of (X_1, X_2) is given by

$$f(x_1, x_2) = \begin{cases} \lambda^2 e^{-\lambda(x_1+x_2)}, \text{ if } x_1 > 0, x_2 > 0 \\ 0, \text{ otherwise} \end{cases}$$

Let $y_1 = \frac{x_1}{x_1+x_2}$ and $y_2 = x_1 + x_2$. Then $x_1 = y_1y_2$ and $x_2 = (1-y_1)y_2$ and

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = y_2 \neq 0 \text{ as } x_1 > 0 \text{ and } x_2 > 0$$

Therefore the joint p.d.f. of $\underline{Y} = (Y_1, Y_2)$ is

$$f_{\underline{Y}}(y_1, y_2) = f(y_1y_2, (1-y_1)y_2)|J|$$

$$= \begin{cases} \lambda^2 y_2 e^{-\lambda y_2}, \text{ if } y_1y_2 > 0, (1-y_1)y_2 > 0 \\ 0, \text{ otherwise} \end{cases}$$

$$= \begin{cases} \lambda^2 y_2 e^{-\lambda y_2}, \text{ if } 0 < y_1 < 1, y_2 > 0 \\ 0, \text{ otherwise} \end{cases}$$

The marginal p.d.f. of Y_1 is

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{\underline{Y}}(y_1, y_2)dy_2$$

$$= \begin{cases} \int_0^{\infty} \lambda^2 y_2 e^{-\lambda y_2} dy_2, \text{ if } 0 < y_1 < 1 \\ 0, \text{ otherwise} \end{cases}$$

$$= \begin{cases} 1, \text{ if } 0 < y_1 < 1 \\ 0, \text{ otherwise} \end{cases}$$

Therefore, $\frac{X_1}{X_1+X_2} \sim U(0,1)$.

Law of Large Numbers, Central Limit Theorem and Normal Approximation

Definition 1. (1) *Let $(X_n)_{n\geq 1}$ be a sequence of random variables (not necessarily independent), and let a be a real number. We say that the sequence $(X_n)_{n\geq 1}$ converges to a in probability (written as $X_n \xrightarrow{p} a$) if for every $\epsilon > 0$, we have*

$$\lim_{n\to\infty} P(\{|X_n - a| \geq \epsilon\}) = 0.$$

(2) *A sequence $(X_n)_{n\geq 1}$ of random variables is said to be bounded in probability if there exists a $M > 0$ such that*

$$P(\cap_{n=1}^{\infty}\{|X_n| \leq M\}) = 1.$$

(3) *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and let F_n be the d.f. of $X_n, n = 1, 2, \ldots$. Let X be a random variable with d.f. F. We say that the sequence $(X_n)_{n\geq 1}$ converges to X in distribution (written as $X_n \xrightarrow{d} X$) if*

$$\lim_{n\to\infty} F_n(x) = F(x), \ \forall\ x \in C_F,$$

where C_F is the set of continuity point of F.

Example 2. *Consider a sequence $(X_n)_{n\geq 1}$ of independent random variables that are uniformly distributed over $(0, 1)$ and let $Y_n = \min\{X_1, \ldots, X_n\}$. The sequence of values of Y_n cannot increase as n increases. Thus, we intuitively expect that Y_n converges to zero.*
Now, for $\epsilon \geq 1$, $P(X_i \geq \epsilon) = 1 - P(X_i \leq \epsilon) = 0$ and for $0 < \epsilon < 1$, $P(X_i \geq \epsilon) = 1 - P(X_i \leq \epsilon) = (1 - \epsilon)$, $1 \leq i \leq n$.

Hence,

$$\begin{aligned} P(\{|Y_n - 0| \geq \epsilon\}) &= P(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon) \\ &= P(X_1 \geq \epsilon) \cdots P(X_n \geq \epsilon) \\ &= \begin{cases} (1-\epsilon)^n, & \text{if } 0 < \epsilon < 1 \\ 0, & \text{if } \epsilon \geq 1 \end{cases} \end{aligned}$$

Hence, for $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(\{|Y_n - 0| \geq \epsilon\}) = 0.$$

Therefore, $Y_n \xrightarrow{p} 0$.

Theorem 3. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables with $E(X_n) = \mu_n$ and $Var(X_n) = \sigma_n^2, n = 1, 2, \ldots$ Suppose $\lim_{n\to\infty} \mu_n = \mu$ and $\lim_{n\to\infty} \sigma_n^2 = 0$. Then $X_n \xrightarrow{p} \mu$.*

Theorem 4. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X be another random variable. Suppose that there exists a $h > 0$ such that m.g.f. $\phi, \phi_1, \phi_2, \ldots$ of $X, X_1, X_2, \ldots$, respectively, are finite on $(-h, h)$.*

(1) *If $\lim_{n\to\infty} \phi_n(t) = \phi(t), \ \forall\ t \in (-h, h)$, then $X_n \xrightarrow{d} X$, where $F, F_1, F_2, \ldots$ are c.d.f. of $X, X_1, X_2, \ldots$, respectively.*
(2) *If $X_1, X_2, \ldots$ are bounded in probability and $X_n \xrightarrow{d} X$, then $\lim_{n\to\infty} \phi_n(t) = \phi(t), \ \forall\ t \in (-h, h)$.*

Continuity Correction: Continuity correction is an adjustment that is made when a discrete distribution is approximated by a continuous distribution.

Table of continuity correction:

Discrete	Continuous
$P(X = a)$	$P(a - 0.5 < X < a + 0.5)$
$P(X > a)$	$P(X > a + 0.5)$
$P(X \leq a)$	$P(X < a + 0.5)$
$P(X < a)$	$P(X < a - 0.5)$
$P(X \geq a)$	$P(X > a - 0.5)$

1. Law of Large Numbers

Theorem 5. ***The weak law of large numbers (WLLN):*** *Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random variables, each having finite mean $E(X_i) = \mu$, $i = 1, 2, \cdots,$. Then, for any $\epsilon > 0$,*

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

i.e.,

$$\lim_{n\to\infty} P\left(\left\{\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\}\right) = 0.$$

equivalently

$$\lim_{n\to\infty} P\left(\left\{\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| < \epsilon\right\}\right) = 1.$$

The weak law of large numbers asserts that the sample mean of a large number of independent identically distributed random variables is very close to the true mean with high probability.

Proof. We assume that the random variables have a finite variance σ^2. Now, $E\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = \frac{1}{n}E(X_1+X_2+\cdots+X_n) = \frac{E(X_1)+E(X_2)+\cdots+E(X_n)}{n} = \mu$, and $Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$.

By Chebyshev's Inequality,

$$P\left(\left\{\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\}\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Since $\lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$, $\lim_{n\to\infty} P\left(\left\{\left|\frac{X_1+X_2+\cdots+X_n}{n} - \mu\right| \geq \epsilon\right\}\right) = 0$. □

Theorem 6. ***The strong law of large numbers (SLLN):*** *Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random variables, each having finite mean $E(X_i) = \mu$, $i = 1, 2, \cdots,$. Then,*

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

i.e.,

$$P\left(\left\{w \in \mathcal{S} \mid \lim_{n\to\infty} \frac{X_1(w) + X_2(w) + \cdots + X_n(w)}{n} = \mu\right\}\right) = 1$$

There is a minor difference between the weak and the strong law. The weak law states that the probability $P\left(\left\{\left|\frac{X_1+X_2+\cdots+X_n}{n} - \mu\right| < \epsilon\right\}\right)$ of a significant deviation of sample mean $\frac{X_1+X_2+\cdots+X_n}{n}$ from μ goes to 1 as $n \to \infty$. Still, for any finite n, this probability can be positive. The weak law provides no conclusive information on the number of such deviations but the strong law does. According to the strong law, $\frac{X_1+X_2+\cdots+X_n}{n}$ converges to μ with probability 1. This implies that for any given $\epsilon > 0$, the probability that the difference $\left|\frac{X_1+X_2+\cdots+X_n}{n} - \mu\right| < \epsilon$ an infinite number of times is equal to 1.

Example 7. *Consider the tossing a coin n-times with S_n the number of heads that turn up. Then the random variable $\frac{S_n}{n}$ represents the fractions of times heads turn up and will have values between 0 and 1. The law of large numbers predicts that the outcomes for this random variable, for large n, will be near $\frac{1}{2}$.*

2. Central Limit Theorem

Theorem 8. ***The Central Limit Theorem (CLT):*** *Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random variables, each having finite mean μ and variance σ^2. Then*

$$Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \xrightarrow{d} Z = N(0,1), \text{ where } \overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

i.e.,

$$Z_n = \frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} Z = N(0,1),$$

i.e.,

$$X_1 + X_2 + \cdots + X_n \approx N(n\mu, n\sigma^2), \text{ for large } n.$$

The Central Limit Theorem states that irrespective of the nature of the parent distribution, the probability distribution of a normalized version of the sample mean, based on a random sample of large size, is approximately standard normal.

Example 9. *Civil engineers believe that W, the amount of weight (in units of 1000 pounds) that a certain span of a bridge can with stand without structural damage resulting, is normally distributed with mean 400 and standard deviation 40. Suppose that the weight (again, in units of 1000 pounds) of a car is a random variable with mean 3 and standard deviation 0.3. How many cars would have to be on the bridge span for the probability of structural damage to exceed 0.1?*

Solution: Let P_n denote the probability of structural damage when there are n cars on the bridge. That is

$$P_n = P(\{X_1 + X_2 + \cdots + X_n \geq W\}) = P(\{X_1 + X_2 + \cdots + X_n - W \geq 0\})$$

where X_i is the weight of the i-th car, $i = 1, 2, \ldots, n$. Now it follows from central limit theorem that $\sum_{i=1}^{n} X_i$ is approximately normal with mean $3n$ and variance $0.09n$. Hence, since W is independent of the $X_i, i = 1, \ldots, n$, and is also normal, it follows that $\sum_{i=1}^{n} X_i - W$ is approximately normal with mean and variance given by

$$E\left(\sum_{i=1}^{n} X_i - W\right) = 3n - 400$$

$$Var\left(\sum_{i=1}^{n} X_i - W\right) = Var\left(\sum_{i=1}^{n} X_i\right) + Var(W) = 0.09n + 1600$$

Therefore, if we let $Z = \frac{\sum_{i=1}^{n} X_i - W - (3n-400)}{\sqrt{0.09n+1600}}$, then

$$P_n = P(\{X_1 + X_2 + \cdots + X_n - W \geq 0\} = P\left(Z \geq \frac{-(3n-400)}{\sqrt{0.09n+1600}}\right)$$

where Y is approximately a standard normal random variable. Now $P(Z \geq 1.28) \approx 0.1$.

$$\{Z \geq 1.28\} \subseteq \{Z \geq \frac{-(3n-400)}{\sqrt{0.09n+1600}}\} \Leftrightarrow 0.1 \leq P_n$$

and

$$\{Z \geq 1.28\} \subseteq \{Z \geq \frac{-(3n-400)}{\sqrt{0.09n+1600}}\} \Leftrightarrow \frac{-(3n-400)}{\sqrt{0.09n+1600}} \leq 1.28$$

or

$$n \geq 117$$

Then there is at least 1 chance in 10 that structural damage will occur.

3. **Normal approximation to Binomial**

Suppose $X \sim \text{Bin}(n,p)$. Then X can be written as $X = X_1 + X_2 + \cdots + X_n$, where

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th trial is success} \\ 0, & \text{otherwise} \end{cases}$$

Also $E(X_i) = p$ and $Var(X_i) = p(1-p), i = 1, 2, \ldots, n$. Therefore, form central limit theorem, the distribution of $\frac{(X_1+X_2+\cdots+X_n)-np}{\sqrt{np(1-p)}}$ approaches the standard normal distribution as $n \to \infty$, i.e., $\frac{X-np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1)$. Hence, X can be approximated with $N(np, np(1-p))$. In general, the normal approximation will be quite good for values of n satisfying $np(1-p) \geq 10$.

Continuity Correction: Since the Normal distribution (it can take all real numbers) is continuous while Binomial distribution is discrete (it can take positive integer values), we should use the integral for Normal distribution with introducing continuity correction so that the discrete integer x in Binomial becomes the interval $(x-0.5, x+0.5)$ in Normal.

Example 10. *A manufacturer makes computer chips of which* 10% *are defective. For a random sample of* 200 *chips, find the approximate probability that more than* 15 *are defective.*

Solution: Let X be the number of defective chips in the sample. Then $X \sim \text{Bin}(200, 0.1)$. therefore, $E(X) = np = 20$ and $Var(X) = np(1-p) = 18$. Then, $\frac{X-20}{\sqrt{18}}$ can be approximated with $Z = N(0,1)$. To allow the continuity correction, we need to calculate $P(X > 15.5)$. So,

$$P(X > 15.5) = P(\frac{X-20}{\sqrt{18}} > \frac{15.5-20}{\sqrt{18}}) = P(Z > -1.06) = P(Z < 1.06) = 0.86.$$

Note: This approximation can also view by using Stirling approximation formula, $n! \approx n^n e^{-n}\sqrt{2\pi n}$, for large n.

4. **Normal approximation to Poisson**

Suppose $X \sim \text{P}(\lambda)$. Then the m.g.f. of X is $M_X(t) = e^{-\lambda(1-e^t)}, \ \forall\ t \in \mathbb{R}$.

Now, let $Y = \frac{X-\lambda}{\sqrt{\lambda}}$. Then the m.g.f. of Y is $M_Y(t) = e^{-t\sqrt{\lambda}} M_X(\frac{t}{\sqrt{\lambda}})$. Therefore,

$$\lim_{\lambda \to \infty} M_Y(t) = e^{\frac{t^2}{2}}$$

This is the m.g.f. of $N(0,1)$. Hence, $\frac{X-\lambda}{\sqrt{\lambda}} \approx N(0,1)$ for large value of λ. In other words, $X \approx N(\lambda, \lambda)$ for large value of λ. If $\lambda \geq 10$, then the normal approximation will be quite good.

Example 11. *Suppose cars arrive at a parking lot at a rate of* 50 *per hour. Assume that the process is a Poisson with* $\lambda = 50$. *Compute the probability that in the next hour number of cars that arrive at this parking lot will be between* 54 *and* 62.

Solution: Let X be the number of cars that arrive at this parking lot. Then

$$P(54 \leq X \leq 62) = \sum_{x=54}^{62} \frac{e^{-50}50^x}{x!}$$

Also, $\frac{X-50}{\sqrt{50}}$ can be approximated with $Z = N(0,1)$. To allow the continuity correction, we need to calculate $P(53.5 \leq X \leq 62.5)$. Now,

$$P(53.5 \leq X \leq 62.5) = P\left(\frac{53.5-50}{\sqrt{50}} \leq \frac{X-50}{\sqrt{50}} \leq \frac{62.5-50}{\sqrt{50}}\right) = \Phi(\frac{12.5}{\sqrt{50}}) - \Phi(\frac{3.5}{\sqrt{50}}) = 0.2717.$$

Point Estimation

Population: In Statistics, population is an aggregate of objects, animate or inanimate, under study. The population may be finite or infinite.
Sample: A part or a finite subset of population is called a sample and the number of units in the sample is called the sample size.
Parameter: The specific characteristics of the population such as population mean (μ), population variance (σ^2) are referred as parameters.
Statistic: It is a function of sample observations, for example, sample mean ($\bar{x}$), sample variance (s^2) are known as statistics.
Here in the theory of point estimation, we consider that the population under study is described by a probability density function (pdf) or probability mass function (pmf), say, $f(x|\underline{\theta})$. The knowledge of parameter(s) $\underline{\theta}$ yields the knowledge of entire population but the problem of statistical parametric inference is that $\underline{\theta}$ is unknown. In order to estimate this $\underline{\theta}$, we resort to take a random sample from the population and infer about the unknown parameter(s) $\underline{\theta}$. It may also happen that instead of $\underline{\theta}$, our interest is to find an estimator for a function of $\underline{\theta}$, say, $g(\underline{\theta})$.

Definition 1. ***Estimator:*** *Any function of the random sample which is used to estimate the unknown value of the given parametric function $g(\underline{\theta})$ is called an estimator. If $\underline{X} = X_1, \ldots, X_n$ is a random sample from a population with common distribution function $F_{\underline{\theta}}$, a function $t(\underline{X})$ used for estimating $g(\underline{\theta})$ is known as an estimator. Let $\underline{x} = x_1, \cdots, x_n$ be a realization of $\underline{X}$. Then, $t(\underline{x})$ is called an estimate.*

For example, in estimating the average height of male students in a class, we may use the sample mean $\bar{X}$ as an estimator. Now, if a random sample of size 20 has a sample mean 170cm, then 170cm is an estimate of the average height of male students of that class.

Parameter Space: The set of all possible values of a parameter(s) is called parameter space. It is denoted by Θ.

Desirable Criteria for Estimators

Given the sample, one may have multiple estimators to estimate the parametric function. For example, to estimate the population average, one may use sample mean/sample median/sample mode. So, in order to choose among the estimators, we should have certain desirable criteria which the estimator to be used should meet. Two such criteria unbiasedness and consistency are discussed as follows.

Definition 2. ***Unbiasedness:*** *Let $X_1, \cdots, X_n$ be a random sample from a population with probability distribution $P_\theta, \theta \in \Theta$. An estimator $t(\underline{X}), \underline{X} = X_1, \cdots, X_n$ is said to be unbiased for estimating $g(\theta)$, if*

$$E_\theta(t(\underline{X})) = g(\theta), \forall \theta \in \Theta. \tag{1}$$

If for some $\theta \in \Theta$, we have

$$E_\theta(t(\underline{X})) = g(\theta) + b_n(\theta),$$

then, $b_n(\theta)$ is called bias of t. If $b_n(\theta) > 0, \forall \theta$, then t is said to overestimate $g(\theta)$. On the other hand if $b_n(\theta) < 0, \forall \theta$, then t is an underestimator of $g(\theta)$.

Definition 3. *An estimator $t(\underline{X})$ is said to be asymptotically unbiased estimator of θ if*

$$\lim_{n\to\infty} b_n(\theta) = 0, \quad \forall \theta \in \Theta. \tag{2}$$

Definition 4. *The quantity $E_\theta(t(\underline{X})-\theta)^2$ is called the mean square error (MSE) of $t(\underline{X})$ about θ.*

$$MSE(t(\underline{X})) = Var(t(\underline{X})) + (b_n(\theta))^2 .$$

If t is unbiased for θ, $MSE(t)$ reduces to $Var(t)$.

Example 5. *Let $X_1, \cdots, X_n$ be a random sample from binomial distribution with parameters n and p, where, n is known and $0 \le p \le 1$. Find unbiased estimators for a) p, the population proportion, b) p^2 c) Variance of X.*

Solution: a) Given that X follows $binomial(n,p)$, n is known and p, the population proportion is unknown. Let $t(\underline{X}) = \frac{X}{n}$, the sample proportion. Now,

$$E(t(\underline{X})) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p.$$

Thus, the sample proportion is an unbiased estimator of population proportion.
b) We can compute that

$$E(X(X-1)) = n(n-1)p^2 \tag{3}$$

Hence, $\frac{X(X-1)}{n(n-1)}$ is an unbiased estimator for p^2.
c)Since, $\text{Var}(X) = np(1-p) = n(p-p^2)$.
Therefore, $t(\underline{X}) = n\left(\frac{X}{n} - \frac{X(X-1)}{n(n-1)}\right) = \frac{X(n-X)}{n-1}$ is an unbiased estimator of Variance of X.

Example 6. *Let $X_1, \cdots, X_n$ be a random sample from the population*

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

.

Is $\bar{X}$ unbiased for θ?

Solution: Note that

$$E(X) = \int_0^\infty x e^{-(x-\theta)} = \theta + 1,$$

so that $E(\bar{X}) = E(X) = \theta + 1$. Thus, $\bar{X}$ is a biased estimator for θ. However, $E(\bar{X} - 1) = \theta$.

Remark 7.

1. *The unbiased estimator need not be unique. For example, let $X_1, \cdots, X_n$ be a random sample form Poisson distribution with parameter $\lambda, \lambda > 0$. Then, $t_1(\underline{X}) = \bar{X}$, $t_2(\underline{X}) = X_i$, $t_3(\underline{X}) = \frac{X_1+2X_2}{3}$ are some unbiased estimators for λ.*
2. *If $E(X)$ exists, then the sample mean is an unbiased estimator of the population mean.*
3. *Let $E(X^2)$ exists, i.e. $Var(X) = \sigma^2$ exists. Then, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is unbiased for σ^2.(Prove!)*
4. *Unbiased estimators may not always exist. For example, X follows binomial distribution with parameters n and p. Then, there exists no unbiased estimator for p^{n+1}.(Prove!)*
5. *Unbiased estimators may not be reasonable always. They may be absurd. For example $t(\underline{X}) = (-2)^X$ is an absurd unbiased estimator for $e^{-3\lambda}$, where, X follows Poisson distribution with parameter λ. (Why?)*

It is intuitively clear that for $t_n(\underline{X})(= t(\underline{X}))$ to be a good estimator the difference $t_n - \theta$ should be as small as possible. However, t_n is a random variable and has its own sampling distribution whose range may be infinitely large. Therefore, it would be sufficient if the sampling distribution of t_n becomes more and more concentrated around θ as the sample size n increases. This means that for each fixed $\theta \in \Theta$, the probability

$$P_\theta[|T_n - \theta| \leq \epsilon]$$

for any given $\epsilon(> 0)$ should be an increasing function of n. This idea leads to the concept of consistency as a criterion of a good estimator.

Definition 8. *Consistency: A statistic t or rather a sequence $\{t_n\}$ is said to be consistent for θ if t_n converges in probability to θ $(t_n \to \theta)$ as $n \to \infty$ for each fixed $\theta \in \Theta$. Thus, t_n is said to be consistent if for every fixed $\theta \in \Theta$ and every pair of positive quantities ϵ and η, however, small, it is possible to find an n_0, depending on ϵ and η, such that*

$$P_\theta[|t_n - \theta| < \epsilon] > 1 - \eta,$$

whenever $n \geq n_0(\epsilon, \eta)$.

If such statistics are used, the accuracy of the estimate increases with the increase in the value of n. It is to be noted that consistency is a large sample property as it is concerned with the behavior of an estimator as the sample size becomes infinitely large.

Example 9. *Let $X_1, \cdots, X_n$ be a random sample from a population with mean μ and variance σ^2. Then,*

$$P(|\bar{X} - \mu| > \epsilon) \leq \frac{Variance(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \quad as \quad n \to \infty.$$

Hence, $\bar{X}$ is consistent for μ.

Example 10. *Let $X_1, \cdots, X_n$ be a sequence of independently and identically distributed (iid) random variables with mean μ, then by weak law of large numbers (WLLN), $\bar{X}$ is consistent for μ.*

Example 11. *Let $\{X_n\}$ be a sequence of iid random variables with pdf*

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

Show that $X_{(1)} = \min X_i$ is a consistent estimator of θ.

Solution: The pdf of $X_{(1)}$ is

$$g(x_{(1)}) = ne^{-(x_{(1)}-\theta)} \left(\int_{x_{(1)}}^{\infty} e^{-(x-\theta)dx} \right)^{n-1} = ne^{-n(x_{(1)})-\theta},$$

for $x_{(1)} > \theta$ and $g(x_{(1)}) = 0$ otherwise.
Now, $P(|X_{(1)} - \theta| < \epsilon) = P(\theta < X_{(1)} < \theta + \epsilon) = 1 - e^{-n\epsilon} \to 1$ as $n \to \infty$.
Hence, $X_{(1)}$ is consistent for θ.

Remark 12. *1. If population mean exists, sample mean is consistent for the population mean.*

2. The consistent estimator may not be unique. For example, if t_n is consistent for θ, then, $\frac{n}{n+1}t_n$, $\frac{n+2}{n+4}t_n$ are all consistent for θ.

Theorem 13. *Let $\{t_n\}$ be a sequence of estimates such that for every $\theta \in \Theta$, the expectation and variance of t_n exist and $E(t_n) = \theta_n \to \theta$ and $V(t_n) \to 0$ as $n \to \infty$. Then, t_n is consistent for θ.*

Proof. We have, by Chebyshev's inequality

$$P[|t_n - \theta| > \epsilon] < \frac{E(t_n - \theta)^2}{\epsilon^2} = \frac{V(t_n) + (E(t_n) - \theta)^2}{\epsilon^2} \to 0 \quad as \quad n \to \infty.$$

□

Theorem 14. *If t is consistent for θ and h is a continuous function of θ. Then, $h(t)$ is consistent for $h(\theta)$.*

Exercise 15. *1. Let $X_1, \ldots, X_n$ be random sample from uniform distribution*

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & otherwise \end{cases}.$$

2. Let $X_1, \ldots, X_n$ be a random sample from the uniform disitribution

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0, & otherwise \end{cases}.$$

Examine the consistencies of the estimators $T_1 = \max X_i$, $T_2 = (n+1) \min X_i$, $T_3 = \min X_i + \max X_i$, $T_4 = 2\bar{X}$ for estimating θ.

Methods of Finding Estimators

There are various methods of finding estimators for the parameters, some of which are listed below.

- Method of Maximum Likelihood
- Method of Moments
- Method of Least Squares
- Method of Minimum Chi square Estimation

We will discuss the method of moments and method of maximum likelihood estimation in detail.

Method of Maximum Likelihood Estimation: Let $X_1, \cdots, X_n$ be a random sample having joint probability density function $f_\theta(x_1, \ldots, x_n)$, $\theta \in \Theta$. The function $f_\theta(x_1, \ldots, x_n)$ may be regarded as a function of θ for given values $(x_1, \ldots, x_n)$. When regarded as a function of θ, the expression $f_\theta(x_1, \ldots, x_n)$ is referred to as the likelihood function of θ, $L(\theta|x_1, \ldots, x_n)$ and expresses the probability that the value of the random variable θ is θ for given values of observations $x_1, \ldots, x_n$. The maximum likelihood estimate (MLE) of θ is that value of θ, within the admissible range of values of θ, which makes the likelihood function a maximum, i.e. the MLE of θ is the number $\hat{\theta}$, if it exists, such that $L(\hat{\theta}|x_1, \ldots, x_n) > L(\theta'|x_1, \ldots, x_n)$ whatever be θ', any other value in Θ.
Ordinarily the parameter θ may be regarded as continuous and in this case the determination of MLE becomes simple. Assuming

1. the likelihood is a positive differentiable function of θ.
2. the maximum of the likelihood does not occur on the boundary of the interval in $\mathbb{R}$ of all admissible values of θ.

The stationary values of the likelihood function within the interval are given by the roots of the equation

$$\frac{\partial L(\theta|x_1, \ldots, x_n)}{\partial \theta} = 0.$$

A sufficient condition that any of these values, say, $\hat{\theta}$ be a real maximum is

$$\frac{\partial^2 L(\theta|x_1, \ldots, x_n)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} < 0.$$

Since $\log L$ attains its maximum value for the same value of θ as L it is usual to maximize $\log L$ in lieu of L. Therefore, we shall seek solution of

$$\frac{\partial \log L(\theta|x_1, \ldots, x_n)}{\partial \theta} = 0. \tag{1}$$

subject to the condition

$$\frac{\partial^2 \log L(\theta|x_1, \ldots, x_n)}{\partial \theta^2} < 0. \tag{2}$$

(2) is generally referred to as likelihood equation. If the observations are iid

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) \tag{3}$$

where $f_\theta(x)$ is the common pdf and here $\log L(\theta) = \sum_{i=1}^{n} \log f_\theta(x)$.

Remark 1. *1. If there are more than one solution satisfying (1) and (2), the maximum of these solutions is to be taken.*

2. We shall ignore any solution which is independent of the observations, i.e., any constant solution.

3. The method holds even if all the variables $X_1, \ldots, X_n$ are discrete and in this case the density function is to be replaced by probability mass function (pmf).

4. If assumptions 1 and 2 do not hold, the MLE cannot be obtained by solving the likelihood equation.

If more than one parameters are involved, i.e., a sample has the pdf $f_\theta(x_1, \ldots, x_n)$ where $\underline{\theta} = (\theta_1, \ldots, \theta_n) \in \Theta \subset \mathbb{R}^k$. In this case, the MLEs are the numbers $\hat{\theta}_1, \ldots, \hat{\theta}_k$, if such a set exists, which maximises f as a function of $\underline{\theta}$. If the likelihood function does not have a maxima on the boundary of set Θ, the maximum of the likelihood function is obtained by the solution of

$$\frac{\partial L(\theta|x_1, \ldots, x_n)}{\partial \theta_i} = 0, \quad i = 1, \ldots, k$$

subject to the condition that the matrix

$$\left(\frac{\partial^2 \log L(\theta|x_1, \ldots, x_n)}{\partial \theta_i \theta_j} \right)_{i,j=1,\ldots,r} \Bigg|_{\underline{\theta}=\underline{\hat{\theta}}} \tag{4}$$

is negative definite.

Example 2. *Let $X_1, \cdots, X_n$ follow Poisson distribution with parameter λ; $\lambda > 0$. Find the MLE for λ.*

Solution: Let $\underline{x} = (x_1, \cdots, x_n)$ be a realization of a random sample. Then the likelihood function is given by

$$L_\lambda(\underline{x}) = \prod_{i=1}^{n} f(x_i, \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}.$$

Therefore, the log likelihood function is given by

$$\log L_\lambda(\underline{x}) = l(\lambda) = -n\lambda + \sum_{i=1}^{n} x_i \log \lambda - \log \left(\prod_{i=1}^{n} x_i! \right).$$

The likelihood equation is

$$\frac{\partial l}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i = 0.$$

Now, $\frac{\sum_{i=1}^n x_i - n\lambda}{\lambda} > 0$ if $\lambda < \bar{x}$ and $\frac{\sum_{i=1}^n x_i - n\lambda}{\lambda} < 0$ if $\lambda > \bar{x}$
Hence, the MLE for λ is $\hat{\lambda} = \bar{x}$.

Example 3. *Let X_1, X_2 be a random sample from a population*

$$f_\theta(x) = \frac{2}{\theta^2}, \quad 0 < x < \theta.$$

Find the MLE of θ.

Solution: The likelihood function is given by

$$L_\theta(\underline{x}) = \frac{4}{\theta^4}(\theta - x_1)(\theta - x^2)$$

The likelihood equation is

$$\frac{\partial \log L}{\partial \theta} = -\frac{4}{\theta} + \frac{1}{\theta - x_1} + \frac{1}{\theta - x_2} = 0.$$

$\Rightarrow$

$$\hat{\theta} = \frac{3(x_1 + x_2) + \sqrt{9(x_1 - x_2)^2 + 4x_1x_2}}{4}.$$

Remark 4. *1. The MLE is unique. (Prove yourself).*

2. ***Invariance Property:*** *If $\hat{\theta}$ is the MLEof θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$ provided $g(\theta)$ is some single valued function of θ.*

Exercise 5. *Let $X_1, \ldots, X_n$ be random sample with following pdf/pmf. Find the MLE(s) of the parameter(s).*

1. $N(\theta, \theta^2)$, $\theta \in (0, \infty)$.
2. $f_{\alpha,\beta}(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}$, $\alpha > 0$, $x \geq \beta > 0$.
3. $P(X_i = 0) = 1 - p$, $P(X_i = 1) = p$ *where* $p \in \left[\frac{1}{4}, \frac{3}{4}\right]$.

Method of Moments: Let $X_1, \cdots, X_n$ be a random sample from a population with probability distribution $P_{\underline{\theta}}; \theta \in \Theta; \underline{\theta} = (\theta_1, \cdots, \theta_k)$.
Consider first k non central moments,

$$\mu_1^{'} = E(X_1) = g_1(\underline{\theta})$$

$$\mu_2^{'} = E(X_1^2) = g_2(\underline{\theta})$$

$$\vdots$$

$$\mu_k^{'} = E(X_1^k) = g_k(\underline{\theta}).$$

Assume that the above system of equations have solution as

$$\theta_1 = h_1(\mu_1^{'}, \cdots, \mu_k^{'})$$

$$\theta_2 = h_2(\mu_1^{'}, \cdots, \mu_k^{'})$$

$$\vdots$$

$$\theta_k = h_k(\mu_1^{'}, \cdots, \mu_k^{'}).$$

Now, define the first k non central sample moments as

$$\alpha_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\alpha_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

$$\vdots$$

$$\alpha_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

In the method of moments, we estimate k^{th} population moment by k^{th} sample moment, i.e.,

$$\hat{\mu}_{j'} = \alpha_j \ ; \quad j = 1, \cdots, k.$$

Thus, the method of moments estimators of $\theta_1, \cdots, \theta_k$ are defined as

$$\hat{\theta}_1 = h_1(\alpha_1, \cdots, \alpha_k)$$

$$\hat{\theta}_2 = h_2(\alpha_1, \cdots, \alpha_k)$$

$$\vdots$$

$$\hat{\theta}_k = h_k(\alpha_1, \cdots, \alpha_k).$$

Example 6. *Let $X_1, \cdots, X_n$ follow $N(\mu, \sigma^2)$; μ and σ^2 are unknown. Find the method of moments estimators μ and σ^2 .*

Solution: We know, for normal distribution, $\mu_1^{'} = \mu$ and $\mu_2{'} = \mu^2 + \sigma^2$. Therefore, we have

$$\mu = \mu_1^{'}$$

and

$$\sigma^2 = \mu_2^{'} - \mu_1^{'\,2}.$$

Now, equating the population moments to sample moments, we get

$$\hat{\mu}_{MME} = \bar{X},$$

and

$$\hat{\sigma}^2_{MME} = \alpha_2 - \alpha_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2.$$

Exercise 7. *1. Let $X_1, \cdots, X_k$ follow binomial distribution with parameters n and p. Find the moment estimators of p, when n in known.*

2. Let $X_1, \cdots, X_n$ be a random sample from Poisson distribution with parameter λ, find the moment estimator of λ.

Remark 8. *1. The method moment estimators need not be unbiased always.*

2. If the functions $g_i's$ are continuous and one-one then the functions $h_i's$ are also continuous and then the method of moment estimators will be consistent.

Testing of Hypotheses

Let $X_1, \ldots, X_n$ be a random sample from a population distribution F_θ, $\theta \in \Theta$, where the functional form of F_θ is known, except, for the parameter θ. For example, we may have a random sample from $N(\mu, 1)$ population where the value μ is unknown. One may be interested in examining the validity of assertion that the value of μ lies in a certain known range, say (μ_1, μ_2) on the basis of the sample drawn from the population. A problem of this type is usually referred to as a problem of testing of hypotheses.

Definition 1. *A parametric hypothesis is a statement about the unknown parameter θ. It is usually referred to as the null hypothesis $H_0 : \theta \in \Theta_0$ where $\Theta_0 \subset \Theta$. The statement $H_1 : \theta \in \Theta_1$ is usually referred to as the alternative hypothesis.*

Our task is to test H_0 against H_1. Here θ can be vector valued also.

Definition 2. *If Θ_0 contains only one point, say θ_0, the hypothesis H_0 is said to be a simple hypothesis. Otherwise, i.e., if Θ_0 contains more than one point, the hypothesis H_0 is said to be composite hypothesis. Similar definition hold for alternative hypothesis.*

Under simple hypothesis the probability density function(pdf) or probability mass function(pmf) of a random variable X is completely specified.

Example 3. *Suppose X follows $N(\mu, \sigma^2)$, where σ^2 is known. The hypothesis $H_0 : \mu = \mu_0$ is a simple hypothesis. The hypotheses $H : \mu > \mu_0$, $H : \mu \leq \mu_0$ are composite hypothesis. If σ^2 is also unknown, $H_0 : \mu = \mu_0$ is composite hypothesis, because, here, $\Theta = \{(\mu, \sigma^2), -\infty < \mu < \infty, \sigma^2 > 0\}$ and $\Theta_0 = \{\mu = \mu_0, \sigma^2 > 0\}$ contains infinitely many points.*

Often we are interested in testing a simple hypothesis $H_0(\theta = \theta_0)$ against alternative composite hypothesis $H_1(\theta \neq \theta_0)$ called two/both sided alternative or one sided composite alternatives $H_1(\theta < \theta_0)$, $H_2(\theta > \theta_0)$.
The problem of testing of hypothesis H_0 against H_1 may be described as follows. Given the sample observations, $\underline{x} = (x_1, \ldots, x_n)'$, we make a decision which will either lead to either acceptance or rejection of H_0. The sample space χ_n of the random vector $\underline{X} = (X_1, \ldots, X_n)'$, is divided into two disjoint subsets w and $w^c = \chi_n - w$ such that H_0 is rejected if $x \in w$ and is accepted if $x \in w^c$. The region w is called the critical region and w^c the region of acceptance. Such a test is called a non-randomized test of H_0 against H_1.
Let $\delta(x)$ is be a function denoting the probability of rejecting the null hypothesis H_0 when $\underline{x}$ is the sample observation. Then, for a non-randomized test

$$\delta(x) = \begin{cases} 1, & if x \in w \\ 0, & if x \in w^c. \end{cases}$$

Definition 4. *Every Borel-measurable mapping $\phi : \mathbb{R}^n \to [0, 1]$ is called a test function.*

Definition 5. *A non-randomized test for testing H_0 against H_1 is a test function $\delta(x)$ defined for all $\underline{x} \in \chi_n$ such that*

$$\delta(x) = \begin{cases} 1, & if x \in w \\ 0, & if x \in w^c. \end{cases}$$

Here, the critical region w depends on the test function δ. If δ changes, $w(\delta)$ will be different. In a non-randomized test if $\underline{x}$ is observed we either accept H_0 or reject it with probability 1.

Definition 6. *A randomised test for testing H_0 against H_1 is a test function $\delta(x)$ defined for all $\underline{x} \in \chi_n$ such that $0 \leq \delta(\underline{x}) \leq 1$ $\forall \underline{x}$. If we observe $\underline{x}$ we make a Bernoulli experiment with probability of success $\delta(\underline{x})$. If a success occurs we reject H_0, otherwise we accept it.*

Randomised test will be needed in general only if X is discrete random variable. But we shall only consider non-randomised tests.

In the problem of testing of hypotheses, the true value of θ remains unknown. We are only aiming at testing whether the observation x supports our assertion $\theta \in \Theta_0$ i.e. $\underline{x}$ is a random sample from the pdf $f_\theta(\underline{x})$, $\theta \in \Theta_0$. Hence, the acceptance (rejection) of H_0 on the basis of $\underline{x}$ does not necessarily imply that H_0 is true (false). Therefore, we may reject H_0 when it is, in fact, true; or we may accept it H_0 when it is false. In both the cases we commit some error.

Definition 7. ***Type I Error:*** *Rejecting H_0, when it is true is known as type I error.*

Definition 8. ***Type II Error:*** *Accepting H_0, when it is false is known as type II error.*

Let $H_0 : \theta = \theta_0$ so that $\Theta_0 = \{\theta_0\}$. In this case probability of type I error is given by

$$P_{\theta_0}(w) = P\{\underline{x} \in w | H_0\} = \int_w f_{\theta_0}(\underline{x}) d\underline{x} = \int \delta(\underline{x}) f_{\theta_0}(\underline{x}) d\underline{x} = E_{\theta_0}(\delta(\underline{x})) \tag{1}$$

and probability of type II error is given by

$$P_\theta(w^c) = P\{\underline{x} \in w^c | H_1\} = \int_w^c f_\theta(\underline{x}) d\underline{x}, \quad \theta \in \Theta_1. \tag{2}$$

The probability of rejecting a true H_0 depends on the test function δ and the value θ_0 and is called the level of significance of test or the size of the critical region w.

The probability of rejecting H_0 when it is false i.e. when $\theta \in \bar{\Theta}_0 = \Theta - \{\theta_0\}$, is

$$P_\theta(x \in w) = 1 - P_\theta(x \in w^c) = 1 - P_\theta(w^c), \quad \theta \in \bar{\Theta}_0 = \gamma_\theta(w), \quad \theta \in \bar{\Theta}_0. \tag{3}$$

$\gamma_\theta(w)$ is called the power of the test w. It depends on the test function δ and the value of the parameter $\theta \in \bar{\Theta}_0$. Note that $\gamma_\theta(w) = E_\theta(\delta(\underline{x}))$ where $\theta \in \bar{\Theta}_0$.

For $\theta \in \Theta_1 (\subseteq (\bar{\Theta}_0)$,

$$\gamma_\theta(w) = 1 - P_\theta(w^c) = 1 - \text{Probability of type II error.} \tag{4}$$

In an ideal test procedure both types of errors should be minimum. However, simultaneous minimization of both both the errors is not possible. Therefore, we try to fix an upper bound on one error and then find a test procedure for which the second probability is minimum. In practice, we pre-assign a small value α (usually 0.05 or 0.01) to probability of type I error and minimize the probability of type II error (β_θ)subject to this constraint.

Example 9. *Let $X_1, \cdots, X_n$ be a random sample from $N(\mu, 1)$. We want to test $H_0 : \mu = -1/2$ against $H_1 : \mu = 1/2$.*
Here, the acceptance region is $w^c = (-\infty, 0]$, i.e., accept H_0 if $\bar{X} \leq 0$. The rejection region is $w = (0, \infty)$, i.e., reject H_0 if $\bar{X} > 0$. Now, we calculate both the errors.

$$\begin{aligned} \alpha = & \quad \textit{Prob(Type I error)= Prob(Rejecting } H_0\textit{, when it is true)} \\ = & \quad P_{\mu=\frac{-1}{2}}(\bar{X} > 0) = P_{\mu=\frac{-1}{2}}(\sqrt{n}(\bar{X} + \tfrac{1}{2}) > \tfrac{\sqrt{n}}{2}) = P(Z > \tfrac{\sqrt{n}}{2}) \\ = & \quad P(Z > 2) = 0.0228, \qquad\qquad \textit{for } n = 16. \end{aligned}$$

β=Prob(Type II error)= Prob(Accepting H_0, when it is false) $\Rightarrow$

$$\beta = P_{\mu=\frac{1}{2}}(\bar{X} \leq 0) = P_{\mu=\frac{-1}{2}}(\sqrt{n}(\bar{X} - \frac{1}{2}) \leq \frac{-\sqrt{n}}{2}) = P(Z \leq \frac{\sqrt{n}}{2}) = P(Z \leq 2) = 0.0228$$

, for $n = 16$.
Here, α and β are same.

Now let us modify the test procedure. Let the acceptance and rejection region be $w_1^c = \{\bar{X} < \frac{-1}{4}\}$ and $w_1 = \{\bar{X} \geq \frac{-1}{4}\}$, respectively. Therefore, the probability of type I and type II errors are

$$\alpha^* = P_{\mu=\frac{-1}{2}}(\bar{X} \geq \frac{-1}{4}) = P_{\mu=\frac{-1}{2}}(\sqrt{n}(\bar{X} + \frac{1}{2}) > \frac{\sqrt{n}}{4}) = P(Z \geq \frac{\sqrt{n}}{4}) = 0.1587,$$

for $n = 16$.

$$\beta^* = P_{\mu=\frac{1}{2}}(\bar{X} < \frac{-1}{4}) = P(Z < -3) = 0.0013,$$

for $n = 16$.
Here, we observe that $\beta^* < \beta$ but $\alpha^* > \alpha$. Hence, it is clear that the simultaneous minimization of both the errors α and β is not possible.

Exercise 10. *Let $X_1, \ldots, X_{20}$ be a random sample from the exponential distribution with pdf $f(x;\theta) = \theta e^{-\theta x}$, $0 < x < \infty$, $\theta > 0$. Calculate type I error and type II error for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$.*

Interval Estimation

In the theory of point estimation, we are interested in estimating the value of parametric function $g(\theta)$ by a single value t based on the observations $x_1, \ldots, x_n$ when the samples are drawn from a density $f(x, \theta)$, $\theta \in Theta$. In practice, one is not generally interested in finding a point estimate of $g(\theta)$, but a set of values, say, $H(\theta)$, such that $H(\theta)$ contains the true value of the parameter $g(\theta)$ with high probability. This type of problems are called problems of confidence interval (set) estimation. When $H(\theta)$ is an interval, it is called confidence interval.

Definition 1. *Confidence Interval: Let $\underline{X} = (X_1, \ldots, X_n)$ be a random sample from a population with density function $f(x, \theta)$, $\theta \in \Theta$. Let $T_1 = t_1(X_1, \ldots, X_n)$, $T_2 = t_2(X_1, \ldots, X_n)$ be two statistics satisfying $T1 \leq T_2$ such that*

$$P_\theta[T_1 \leq g(\theta) \leq T_2] = 1 - \alpha \quad \forall \theta \in \Theta \tag{1}$$

where $(1-\alpha)$ does not depend upon θ. Then the random interval (T_1, T_2) is called the $100(1-\alpha)\%$ confidence interval for $g(\theta)$. The quantity $(1-\alpha)$ is called the confidence coefficient of this interval. The statistics T_1, T_2 are respectively called lower and upper confidence limits for $g(\theta)$. For a given sample observation $\underline{x} = (x_1, \ldots, x_n)$, the values of the statistic $T_1(\underline{x})$, $T_2(\underline{x})$ are the confidence limits for $g(\theta)$.

Usually α is taken to be very small quantity, 0.05, 0.01(say) so that $1-\alpha$ is 0.95, 0.99. In some cases, any of the two statistics T_1, T_2 may be a constant; however $T1$, T_2 can not both be constants.

Definition 2. *One-sided confidence interval: Let X be a random sample from a population with pdf $f(x, \theta)$, $\theta \in \Theta$. Let $T_1 = t_1(\underline{X})$ be a statistic such that*

$$P_\theta[T_1 \leq g(\theta)] = 1 - \alpha \quad \theta \in \Theta \tag{2}$$

where α does not depend upon θ. Then T_1 is called the on-sided lower confidence limit for $g(\theta)$. Thus, the interval (T_1, ∞) covers $g(\theta)$ with probability $1-\alpha$. Similarly, let $T_2 = t_2(\underline{X})$ be a statistic such that

$$P_\theta[T_2 \geq g(\theta)] = 1 - \alpha \quad \theta \in \Theta \tag{3}$$

where α does not depend on θ. Then T_2 is called the one-sided upper confidence limit for $g(\theta)$. Here, the interval $(-\infty, T_2)$ covers $g(\theta)$ with probability $(1-\alpha)$. Note that θ may be a vector of parameters. In making probability statements like (1), (2) and (3), we do not mean that $g(\theta)$ is a random variable. (1) means that the probability is $(1-\alpha)$ that the random interval (T_1, T_2) will cover $g(\theta)$, where the true value of the parameter θ may be.

Example 3. *Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ population when σ^2 is known. Find a $100(1-\alpha)\%$ confidence interval for μ.*

Solution: It is known that

$$P_\mu\left(\bar{X} - \tau_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad \forall \mu, \tag{4}$$

where, $\tau_{\alpha/2}$ is the upper $100(\alpha/2)\%$ probability point on a standard normal distribution.
Hence the interval $\left(\bar{X} - \tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + \tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ is $100(1-\alpha\%)$ confidence interval for μ.

Clearly any random interval $\left(\bar{X} - \tau_{\alpha_1}\frac{\sigma}{\sqrt{n}}, \bar{X} + \tau_{\alpha_2}\frac{\sigma}{\sqrt{n}}\right)$ where $\alpha = \alpha_1 + \alpha_2$ is $100(1-\alpha\%)$ confidence interval for μ. Again

$$P_\mu\left(\bar{X} - \tau_\alpha\frac{\sigma}{\sqrt{n}} \leq \mu\right) = 1 - \alpha \quad \forall \mu. \tag{5}$$

Also

$$P_\mu\left(\bar{X}+\tau_\alpha\frac{\sigma}{\sqrt{n}}\geq\mu\right)=1-\alpha\quad\forall\mu. \tag{6}$$

Therefore $T_1=\bar{X}-\tau_\alpha\frac{\sigma}{\sqrt{n}}$, $T_2=\bar{X}+\tau_\alpha\frac{\sigma}{\sqrt{n}}$, are respectively the lower and upper confidence limits for μ.

In case of discrete random variables, it is evident that it is not possible to construct confidence intervals of exact confidence coefficient $(1-\alpha)$ for each $0<\alpha<1$. In this case, we may construct confidence intervals of confidence coefficient measuring at least $(1-\alpha)$. The statistics (T_1,T_2) will, therefore, provide confidence limits to a parameter $g(\theta)$ if

$$P_\theta(T_1\leq g(\theta))\leq T_2)\geq 1-\alpha\quad\theta\in\Theta. \tag{7}$$

Similarly, $T_1(T_2)$ will be lower(upper) confidence limit with confidence coefficient $(1-\alpha)$ if $P_\theta(T_1\leq g(\theta))\geq 1-\alpha$ $(P_\theta(T_2\geq g(\theta))\geq 1-\alpha)\ \forall\ \theta\in\Theta$.

Linear Regression

Regression analysis is a mathematical measure of average relationship between two or more variables in terms of the original units of data. The curve around which the scatter diagram of the two variables cluster around is called the curve of regression. If this curve looks like a straight line, it is known as line of regression.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. This line of best fit is obtained by the principle of least squares. Let $(x_i, y_i); i = 1, 2, \cdots, n$ be the given data, where Y is the dependent variable and X is the independent variable. Let the line of regression of Y on X be

$$Y = a + bX. \tag{1}$$

Here, (1) represents a family of straight lines for different values of the arbitrary constants a and b. The problem is to determine the values of a and b so that the line (1) gives the best fit to the given data. For this, we use the principle of least squares. Let (x_i, y_i) be an observation of the given data. As per our assumption, we have $y_i = a + bx_i$. Hence, the error of the estimate or the residue for y_i is given by

$$\epsilon_i = y_i - a - bx_i.$$

Therefore, the total error sum of squares is given by

$$E = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{2}$$

According to the principle of least squares, we have to determine a and b so that the total error sum of squares E is minimum. From the principle of maxima and minima, the partial derivatives of E with respect to a and b are given should vanish separately, i.e.,

$$\frac{\partial E}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0,$$

which gives

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i. \tag{3}$$

$$\frac{\partial E}{\partial b} = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0,$$

which gives

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2. \tag{4}$$

(5) and (6) are known as normal equations for estimating a and b. All the quantities in (5) and (6) can be obtained form the given set of data except a and b and hence these equations can be solved for a and b. With these values of a and b so obtained, (1) is the line of best fit for the given data set $(x_i, y_i); i = 1, 2, \cdots, n$.

Now, dividing (3) by n we get

$$\bar{y} = a + b\bar{x} \tag{5}$$

Thus, the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$. Therefore, (4) implies

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = b\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

which implies

$$b = b_{yx} = \frac{cov(X,Y)}{var(X)}. \tag{6}$$

Here, b_{yx} is the coefficient of regression of Y on X and it is also the slope of the line of regression Y on X. Since the line of regression passes through $(\bar{x}, \bar{y})$, its equation can also be written as

$$Y - \bar{y} = b_{yx}(X - \bar{x}) = \frac{cov(X,Y)}{var(X)}(X - \bar{x}),$$

which gives the line of regression Y on X as

$$Y - \bar{y} = \rho\frac{\sigma_y}{\sigma_x}(X - \bar{x}). \tag{7}$$

where ρ is the Karl Pearson's coefficient of correlation

Exercise 1. *1. Derive the expression for line of regression X on Y.*

2. Obtain the equations of two lines of regression for the following data. Also obtain the estimate of X for $Y = 70$.

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71