Statistical Inference and Multivariate Analysis (MA 324)

Class Notes
January – May, 2021

Instructor
Ayon Ganguly
Department of Mathematics
IIT Guwahati

Contents

Chapter 1

Review

In this chapter, we will recall some of the concepts and theorems that were covered in the course Probability Theory and Random Processes (MA 225). We will use these concepts and theorems in this course.

1.1 Transformation Techniques

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

1.1.1 Technique 1

In Technique 1, we try to find the joint cumulative distribution function (JCDF) of $\boldsymbol{Y} = g(\boldsymbol{X})$ given the distribution of $\boldsymbol{X}$. Recall that, for a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, the JCDF at $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is $P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$. As the JCDF exists for any random vector, this technique, in principle, can be used for any random vector. This technique is best understood using examples.

Example 1.1. Let the random variable (RV) X has the following probability mass function (PMF):

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

Consider $Y = X^2$. Let us denote the cumulative distribution function (CDF) of a random variable X by $F_X(\cdot)$. Clearly, for $y < 0$, $F_Y(y) = P(X^2 \leq y) = P(X \in \emptyset) = 0$. For $y \geq 0$,

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

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

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

For $1 \leq y < 4$,

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

For $4 \leq y < 9$,

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

For $y \geq 9$,

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

Hence, the CDF of Y is

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

Example 1.2. Let the RV X has the following probability density function (PDF):

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

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

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

For $1 \leq y < 4$,

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

For $y \geq 4$,

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

Example 1.3. Let the RV X has the following PDF:

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

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

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

For $1 \leq y < 2$,

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

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

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

Thus, the CDF of Y is given by

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

Example 1.4. Let X_1 and X_2 be independent and identically distributed (*i.i.d.*) $U(0, 1)$ random variables. Suppose we want to find the CDF of $Y = X_1 + X_2$. Now,

$$F_Y(y) = P(Y \leq y) = P(X_1 + X_2 \leq y) = \int\int_{x_1+x_2 \leq y} f_{X_1, X_2}(x_1, x_2)dx_1dx_2. \tag{1.1}$$

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

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

Thus, the JPDF of (X_1, X_2) is positive only on the unit square $(0, 1) \times (0, 1)$, which is indicated by gray shade in Figure 1.1. Now, to compute the integration in (1.1), we need to consider the following cases.

For $y < 0$, consider the Figure 1.1a. As the integrand in (1.1) is zero over the region $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$ for $y < 0$,

$$F_Y(y) = 0.$$

For $0 \leq y < 1$, consider the Figure 1.1b. The integrand is positive only on the shaded region in the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$. Therefore,

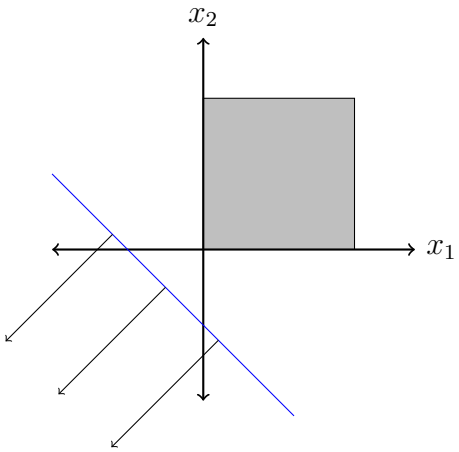
$$F_Y(y) = \int_{0}^{y} \int_{0}^{y-x_2} dx_1dx_2 = \frac{1}{2}y^2.$$

For $1 \leq y < 2$, consider the Figure 1.1c. The integrand is positive only on the shaded region in the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$. Therefore,

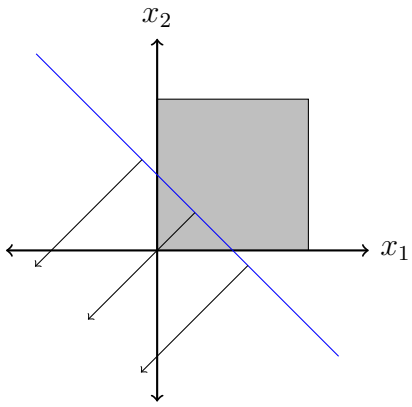
$$F_Y(y) = 1 - \int_{y-1}^{1} \int_{y-x_2}^{1} dx_1dx_2 = 1 - \frac{1}{2}(2 - y)^2.$$

For $y \geq 2$, consider the Figure 1.1d. The integrand is positive on the shaded region in the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$ and the square $(0, 1) \times (0, 1)$ is completely inside the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$. Therefore,

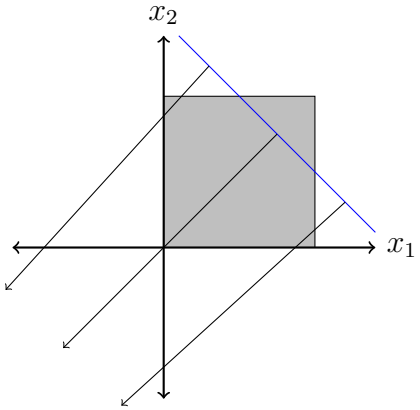
$$F_Y(y) = 1.$$



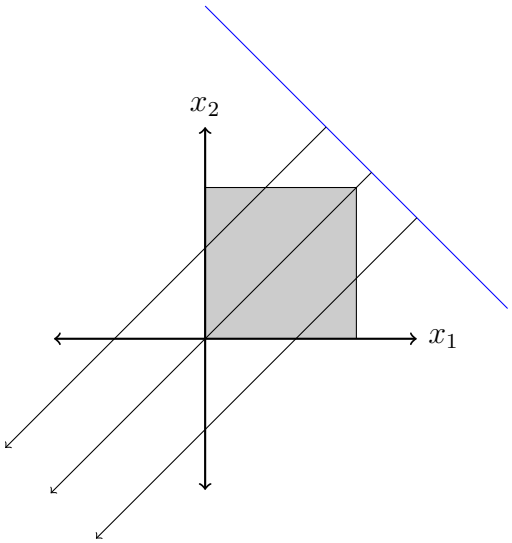
(a) $y < 0$



(b) $0 \leq y < 1$



(c) $1 \leq y < 2$



(d) $y \geq 2$

Figure 1.1: Plot for Example 1.4

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

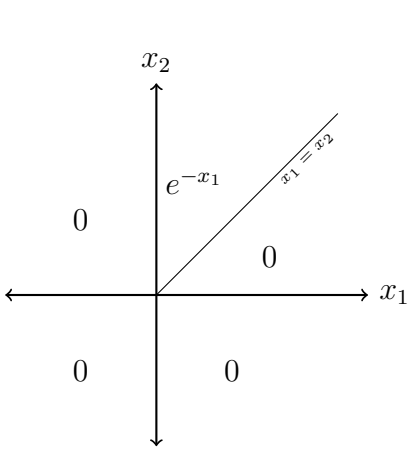
$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2}y^2 & \text{if } 0 \leq y < 1 \\ 1 - \frac{1}{2}(2-y)^2 & \text{if } 1 \leq y < 2 \\ 1 & \text{if } y \geq 2. \end{cases} \qquad ||$$

Example 1.5. Let the JPDF of (X_1, X_2) be given by

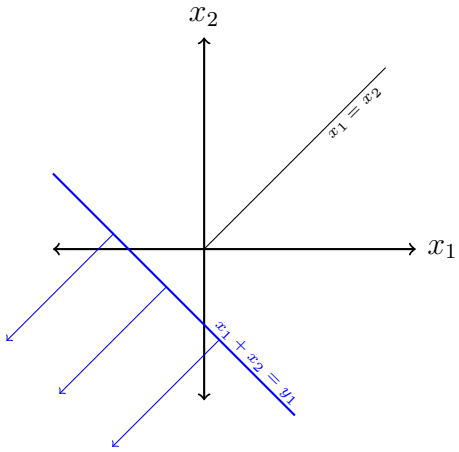
$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} e^{-x_1} & \text{if } 0 < x_1 < x_2 < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that we want to find the JCDF of $Y_1 = X_1 + X_2$ and $Y_2 = X_2 - X_1$. Note that the JPDF of (X_1, X_2) is positive only on the set $S_{X_1, X_2} = \{(x_1, x_2) \in \mathbb{R}^2 : 0 < x_1 < x_2 < \infty\}$. See Figure 1.2a. Now, let $A_{y_1, y_2} = \{(x_1, x_2) \in \mathbb{R} : x_1 + x_2 \leq y_1, x_2 - x_1 \leq y_2\}$. Then

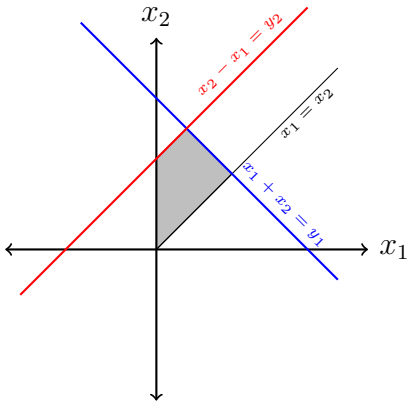
$$F_{Y_1, Y_2}(y_1, y_2) = P(X_1 + X_2 \leq y_1, X_2 - X_1 \leq y_2) = \int\int_{A_{y_1, y_2}} f_{X_1, X_2}(x_1, x_2) dx_2 dx_1. \quad (1.2)$$



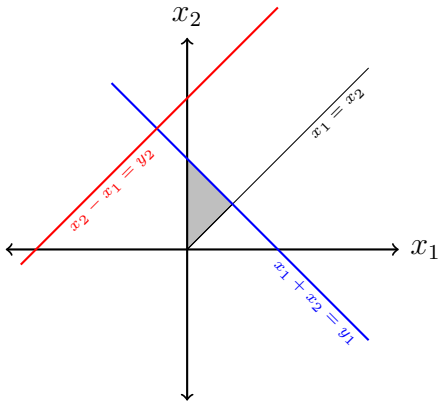
(a) JPDF of (X_1, X_2)



(b) $y_1 < 0$



(c) $0 < y_2 < y_1 < \infty$



(d) $0 < y_1 < y_2 < \infty$

Figure 1.2: Plot for Example 1.5

Suppose that $y_1 < 0$. Then $F_{Y_1}(y_1) = 0$. See the Figure 1.2b. As $F_{Y_1,Y_2}(y_1, y_2) \leq \min\{F_{Y_1}(y_1), F_{Y_2}(y_2)\}$, $F_{Y_1,Y_2}(y_1, y_2) = 0$ for $y_1 < 0$. Similarly, $F_{Y_1,Y_2}(y_1, y_2) = 0$ for $y_2 < 0$. For $0 < y_2 < y_1 < \infty$, $A_{y_1,y_2} \cap S_{X_1,X_2}$ is the shaded region of the Figure 1.2c. Therefore,

$$\begin{aligned} F_{Y_1,Y_2}(y_1, y_2) &= \int_0^{\frac{y_1-y_2}{2}} \int_{x_1}^{x_1+y_1} e^{-x_1} dx_2 dx_1 + \int_{\frac{y_1-y_2}{2}}^{\frac{y_1}{2}} \int_{x_1}^{y_1-x_1} e^{-x_1} dx_2 dx_1 \\ &= y_1 + e^{-\frac{y_1}{2}} - (y_1 - y_2 + 2)e^{-\frac{y_1-y_2}{2}}. \end{aligned}$$

For $0 < y_1 < y_2 < \infty$, $A_{y_1,y_2} \cap S_{X_1,X_2}$ is indicated by the shaded region in the Figure 1.2d. Therefore,

$$F_{Y_1,Y_2}(y_1, y_2) = \int_0^{\frac{y_1}{2}} \int_{x_1}^{y_1-x_1} e^{-x_1} dx_2 dx_1 = y_1 + 2e^{-\frac{y_1}{2}} - 2.$$

Thus, the JCDF of $(Y_1, Y_2) = (X_1 + X_2, X_2 - X_1)$ is given by

$$F_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 0 & \text{if } y_1 < 0 \text{ or } y_2 < 0 \\ y_1 + e^{-\frac{y_1}{2}} - (y_1 - y_2 + 2)e^{-\frac{y_1-y_2}{2}} & \text{if } 0 < y_2 \leq y_1 < \infty \\ y_1 + 2e^{-\frac{y_1}{2}} - 2 & \text{if } 0 < y_1 < y_2 < \infty. \end{cases} \quad ||$$

The basic idea here is to write the event $\boldsymbol{Y} \leq \boldsymbol{y}$ as $\boldsymbol{X} \in A_{\boldsymbol{y}}$ for appropriate set $A_{\boldsymbol{y}}$. In Example 1.3, we have written $Y \leq y$ as $X \in (-\infty, i+1)$ for $y \in [i,\, i+1)$. Then using the distribution of X, one needs to find the probability of the event $X \in A_{\boldsymbol{y}}$.

1.1.2 Technique 2

In Technique 2, we try to find joint probability mass function (JPMF) (if $\boldsymbol{Y}$ isa discrete random vector) or JPDF (if $\boldsymbol{Y}$ is a continuous random vector) of Y directly without finding its' CDF. Obviously, first we need to understand whether $\boldsymbol{Y}$ is discrete random vector or continuous random vector. This technique is mainly based on two theorems. The first theorem consider the case when $\boldsymbol{X}$ is discrete random vector. We will see that if $\boldsymbol{X}$ is discrete random vector, then $\boldsymbol{Y}$ is also a discrete random vector. The second theorem addresses the case when $\boldsymbol{X}$ is continuous random vector. We will see the under some condition, $\boldsymbol{Y}$ is a continuous random vector if $\boldsymbol{X}$ is continuous random vector. With examples, we will illustrate that if the conditions do not hold, then $\boldsymbol{Y}$ can be discrete random vector as well as continuous random vector. Hence, those conditions are important.

Theorem 1.1. *Let $\boldsymbol{X} = (X_1,\, X_2,\, \ldots,\, X_n)$ be a discrete random vector with joint probability mass function (JPMF) $f_{\boldsymbol{X}}$ and support $S_{\boldsymbol{X}}$. Let $g_i : \mathbb{R}^n \to \mathbb{R}$ for all $i = 1,\, 2,\, \ldots,\, k$. Let $Y_i = g_i(\boldsymbol{X})$ for $i = 1,\, 2,\, \ldots,\, k$. Then $\boldsymbol{Y} = (Y_1, \ldots,\, Y_k)$ is a discrete random vector with JPMF*

$$f_{\boldsymbol{Y}}(y_1,\, \ldots,\, y_k) = \begin{cases} \displaystyle\sum_{\boldsymbol{x} \in A_{\boldsymbol{y}}} f_{\boldsymbol{X}}(\boldsymbol{x}) & \text{if } (y_1,\, \ldots,\, y_k) \in S_{\boldsymbol{Y}} \\ 0 & \text{otherwise,} \end{cases}$$

where $A_{\boldsymbol{y}} = \{\boldsymbol{x} \in S_{\boldsymbol{X}} : g_i(\boldsymbol{x}) = y_i,\, i = 1,\, \ldots,\, k\}$ and $S_{\boldsymbol{Y}} = \{(g_1(\boldsymbol{x}),\, \ldots,\, g_k(\boldsymbol{x})) : \boldsymbol{x} \in S_{\boldsymbol{X}}\}$.

Proof: The proof of the theorem is skipped. □

Example 1.6. Let the RV X has the following PMF:

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

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

$$P\,(Y = 0) = P\,(X = 0) = \frac{1}{7}.$$
$$P\,(Y = 1) = P\,(X = 1 \text{ or } -1) = \frac{2}{7}.$$
$$P\,(Y = 4) = P\,(X = 2 \text{ or } -2) = \frac{5}{14}.$$
$$P\,(Y = 9) = P\,(X = 3 \text{ or } -3) = \frac{3}{14}.$$

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

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

Example 1.7. Let $X \sim Bin(n, p)$. Suppose that we are interested to find the distribution of $Y = n - X$. As X is a DRV, using the above theorem, Y is also DRV. Here, $S_X = \{0, 1, \ldots, n\} = S_Y$. For any $y \in S_Y$, $A_y = \{n - y\}$. Hence, the PMF of Y is

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

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

Example 1.8. Let $X_1 \sim Poi(\lambda_1)$ and $X_2 \sim Poi(\lambda_2)$. Also, assume that X_1 and X_2 are independent. Then $Y = X_1 + X_2 \sim Poi(\lambda_1 + \lambda_2)$. To see it, we can apply Theorem 1.1. First note that the JPMF of (X_1, X_2) is given by

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} \frac{e^{-(\lambda_1+\lambda_2)} \lambda_1^{x_1} \lambda_2^{x_2}}{x_1! x_2!} & \text{if } x_1 = 0, 1, \ldots; x_2 = 0, 1, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $S_{X_1, X_2} = \{0, 1, 2, \ldots\} \times \{0, 1, 2, \ldots\}$, which implies that $S_Y = \{0, 1, 2, \ldots\}$. For $y \in S_Y$, $A_y = \{(x, y-x) : x = 0, 1, \ldots, y\}$. Hence, using the Theorem 1.1, for $y \in S_Y$,

$$f_Y(y) = \sum_{(x_1, x_2) \in A_y} \frac{e^{-(\lambda_1+\lambda_2)} \lambda_1^{x_1} \lambda_2^{x_2}}{x_1! x_2!} = \frac{e^{-(\lambda_1+\lambda_2)}}{y!} \sum_{x=0}^{y} \binom{y}{x} \lambda_1^x \lambda_2^{y-x} = \frac{1}{y!} e^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^y .$$

Thus, the PMF of $Y = X_1 + X_2$ is

$$f_Y(y) = \begin{cases} \frac{1}{y!} e^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^y & \text{if } y = 0, 1, \ldots \\ 0 & \text{otherwise,} \end{cases}$$

which is PMF of a $P(\lambda_1 + \lambda_2)$. Hence, $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$. ||

Example 1.9. Let $X_1 \sim Bin(n_1,\, p)$ and $X_2 \sim Bin(n_2,\, p)$. We also assume that X_1 and X_2 are independent. Suppose that we want to find the PMF of $Y = X_1 + X_2$. Note that X_1 and X_2 are the numbers of successes out of n_1 and n_2 independent Bernoulli trials, respectively. In both the cases the probability of success is p. Therefore, Y is the number of successes out of $n_1 + n_2$ Bernoulli trials with success probability p. As X_1 and X_2 are independent, these $n_1 + n_2$ Bernoulli trials can be assumed to be independent. Hence, the distribution of Y must be $Bin(n_1 + n_2,\, p)$. Let us now check if we get the same distribution using the Theorem 1.1. The JPMF of X_1 and X_2 is

$$f_{X_1,\,X_2}(x_1,\, x_2) = \begin{cases} \binom{n_1}{x_1}\binom{n_2}{x_2} p^{x_1+x_2}(1-p)^{n_1+n_2-x_1-x_2} & \text{if } x_1 = 0,\, 1,\, \ldots,\, n_1;\ x_2 = 0,\, 1,\, \ldots,\, n_2 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $S_{X_1,\,X_2} = \{0,\, 1,\, \ldots,\, n_1\} \times \{0,\, 1,\, \ldots,\, n_2\}$. Without loss of generality, we assume that $n_1 \leq n_2$. If not, exchange the roles of X_1 and X_2. Now, $S_Y = \{0,\, 1,\, \ldots,\, n_1 + n_2\}$. For $y \in S_Y$,

$$\begin{aligned} A_y &= \{(x_1,\, x_2) \in S_{X_1,\,X_2} : x_1 + x_2 = y\} \\ &= \begin{cases} \{(x,\, y-x) : x = 0,\, 1,\, \ldots,\, y\} & \text{if } 0 \leq y \leq n_1 \\ \{(x,\, y-x) : x = 0,\, 1,\, \ldots,\, n_1\} & \text{if } n_1 < y \leq n_2 \\ \{(x,\, y-x) : x = y - n_2,\, \ldots,\, n_1\} & \text{if } n_2 < y \leq n_1 + n_2. \end{cases} \end{aligned}$$

Hence, for $y \in S_Y$ and $y \leq n_1$,

$$f_Y(y) = \sum_{x=0}^{y} \binom{n_1}{x}\binom{n_2}{y-x} p^y (1-p)^{n_1+n_2-y} = \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y}.$$

The last equality can be proved by collecting the coefficient of x^y from both sides of the following expression:

$$(1+x)^{n_1}(1+x)^{n_2} = \left\{\sum_{i=0}^{n_1}\binom{n_1}{i}x^i\right\} \times \left\{\sum_{i=0}^{n_2}\binom{n_2}{i}x^i\right\}.$$

For $y \in S_Y$ and $n_1 < y \leq n_2$,

$$f_Y(y) = \sum_{x=0}^{n_1} \binom{n_1}{x}\binom{n_2}{y-x} p^y (1-p)^{n_1+n_2-y} = \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y}.$$

For $y \in S_Y$ and $n_2 < y \leq n_1 + n_2$,

$$f_Y(y) = \sum_{x=y-n_2}^{n_1} \binom{n_1}{x}\binom{n_2}{y-x} p^y (1-p)^{n_1+n_2-y} = \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y}.$$

Thus, $X_1 + X_2 \sim Bin\,(n_1 + n_2,\, p)$. Note that independence of X_1 and X_2 and same value of probability of success are important for the result. ||

Theorem 1.2. *Let $\boldsymbol{X} = (X_1,\, \ldots,\, X_n)$ be a continuous random vector with JPDF $f_{\boldsymbol{X}}$.*

1. *Let $y_i = g_i(\boldsymbol{x})$, $i = 1,\, 2,\, \ldots,\, n$ be $\mathbb{R}^n \to \mathbb{R}$ functions such that*

$$\boldsymbol{y} = g(\boldsymbol{x}) = (g_1(\boldsymbol{x}),\, \ldots,\, g_n(\boldsymbol{x}))$$

is one-to-one. That means that there exists the inverse transformation $x_i = h_i(\boldsymbol{y})$, $i = 1,\, 2,\, \ldots,\, n$ defined on the range of the transformation.

2. *Assume that both the mapping and its' inverse are continuous.*

3. *Assume that partial derivatives* $\frac{\partial x_i}{\partial y_j}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$, *exist and are continuous.*

4. *Assume that the Jacobian of the inverse transformation*

$$J \doteq \det\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j=1,2,\ldots,n} \neq 0$$

on the range of the transformation.

Then $\boldsymbol{Y} = (g_1(\boldsymbol{X}), \ldots, g_n(\boldsymbol{X}))$ *is a continuous random vector with JPDF*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(h_1(\boldsymbol{y}), \ldots, h_n(\boldsymbol{y}))|J|.$$

Proof: The proof of this theorem can be done using transformation of variable technique for multiple integration. However, the proof is skipped here. □

Remark 1.1. Note that g is a vector valued function. As g should be one-to-one, the dimension of g should be same as dimension of the argument of g. Though we have written that $g_i : \mathbb{R}^n \to \mathbb{R}$ in the previous theorem, the conclusion of the theorem is valid if we replace $g_i : \mathbb{R}^n \to \mathbb{R}$ by $g_i : S_{\boldsymbol{X}} \to \mathbb{R}$. Moreover, the theorem gives us sufficient conditions for $g(\boldsymbol{X})$ to be a continuous random vector, when $\boldsymbol{X}$ is continuous random vector. Thus, $g(\boldsymbol{X})$ can be a continuous random vector even if the conditions of the previous theorem do not hold true. †

Example 1.10. Let $X \sim U(0, 1)$. Suppose that $g(x) = -\ln x$ for $x \in (0, 1)$. Also, the support of X is $S_X = (0, 1)$, which is an interval. Clearly, $g'(x) < 0$ for all $x \in (0, 1)$. The inverse of $g(\cdot)$ is $g^{-1}(y) = e^{-y}$ for all $y \in g(S_X) = (0, \infty)$. Hence, $Y = -\ln X$ is a continuous random variable (CRV) with PDF

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

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

Example 1.11. Let $X \sim Exp(1)$. Suppose we are interested to find the distribution of $Y = X^2$. Here, $g(x) = x^2$ for $x \in S_X = (0, \infty)$. Also, $g'(x) = 2x > 0$ for all $x > 0$. Hence, $Y = X^2$ is a CRV. Note that $g^{-1}(y) = \sqrt{y}$. Thus, the PDF of Y is

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

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

Example 1.12. Let $X \sim N(0, 1)$. Suppose that we want to find the distribution of $Y = X^2$. Note that the support of X is $\mathbb{R}$ and $g'(x) = 2x$ does not take only positive or negative values on $\mathbb{R}$. Hence, we cannot use Theorem 1.2. However, we can use technique 1 to obtain the CDF of Y and then check the type of the RV Y. The CDF of Y is given by

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

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

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

Thus, Y is a CRV. This example shows that even if some of the conditions of the Theorem 1.2 do not hold true, the RV Y could be CRV. Thus, the conditions in the Theorem 1.2 are important and they are sufficient conditions, but not necessary. ||

Example 1.13. Let X_1 and X_2 be *i.i.d.* $U(0, 1)$ random variables. We want to find the JPDF of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Clearly,

$$g_1(x_1, x_2) = x_1 + x_2 \quad \text{and} \quad g_2(x_1, x_2) = x_1 - x_2.$$

Thus, $\boldsymbol{y} = (y_1, y_2) = g(x_1, x_2) = (g_1(x_1, x_2), g_2(x_1, x_2)) = (x_1 + x_2, x_1 - x_2)$. Now, if $(x_1, x_2) \neq (\tilde{x}_1, \tilde{x}_2)$, then $g(x_1, x_2) \neq g(\tilde{x}_1, \tilde{x}_2)$. If not, then $x_1 + x_2 = \tilde{x}_1 + \tilde{x}_2$ and $x_1 - x_2 = \tilde{x}_1 - \tilde{x}_2$, which implies $x_1 = \tilde{x}_1$ and $x_2 = \tilde{x}_2$. This is a contradiction. Hence, the function $g(\cdot, \cdot)$ is one-to-one. The inverse function is given by $h(y_1, y_2) = (h_1(y_1, y_2), h_2(y_1, y_2))$, where $x_1 = h_1(y_1, y_2) = \frac{1}{2}(y_1 + y_2)$ and $x_2 = h_2(y_1, y_2) = \frac{1}{2}(y_1 - y_2)$. Clearly, both the mapping and inverse mapping are continuous. Now,

$$\frac{\partial x_1}{\partial y_1} = \frac{1}{2}, \frac{\partial x_1}{\partial y_2} = \frac{1}{2}, \frac{\partial x_2}{\partial y_1} = \frac{1}{2}, \quad \text{and} \quad \frac{\partial x_2}{\partial y_2} = -\frac{1}{2}.$$

All the partial derivatives are continuous. The Jacobian is

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2} \neq 0.$$

Thus, all the four conditions of the Theorem 1.2 hold, and hence, $\boldsymbol{Y} = (Y_1, Y_2)$ is a continuous random vector with JPDF

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= f_{X_1, X_2}\left(\frac{1}{2}(y_1 + y_2), \frac{1}{2}(y_1 - y_2)\right)\left|-\frac{1}{2}\right| \\ &= \begin{cases} \frac{1}{2} & \text{if } 0 < y_1 + y_2 < 2,\ 0 < y_1 - y_2 < 2 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Note that in Example 1.4, we have found the distribution of $X_1 + X_2$. You may find the marginal distribution of $X_1 + X_2$ from JPDF above and check if you are getting same marginal distribution. ||

Example 1.14. Let X_1 and X_2 be *i.i.d.* $N(0,\,1)$ random variables. We want to find the PDF of $Y_1 = X_1/X_2$. Note that we cannot use Theorem 1.2 directly here as we have a single function $g_1(x_1,\,x_2) = \frac{x_1}{x_2}$. Thus, we need to bring an auxiliary new function $g_2(x_1,\,x_2)$ such that $g(x_1,\,x_2) = (g_1(x_1,\,x_2),\,g_2(x_1,\,x_2))$ satisfies all the conditions of Theorem 1.2. Let us take $g_2(x_1,\,x_2) = x_2$. Clearly, $g(x_1,\,x_2)$ is a one-to-one function. Here, the inverse function is $h(y_1,\,y_2) = (h_1(y_1,\,y_2),\,h_2(y_1,\,y_2))$, where $x_1 = h_1(y_1,\,y_2) = y_1y_2$ and $x_2 = h_2(y_1,\,y_2) = y_2$. It is easy to see that mapping g and its' inverse are continuous. Also,

$$\frac{\partial x_1}{\partial y_1} = y_2,\ \frac{\partial x_1}{\partial y_2} = y_1,\ \frac{\partial x_2}{\partial y_1} = 0, \quad \text{and} \quad \frac{\partial x_2}{\partial y_2} = 1.$$

All the partial derivatives are continuous. Hence, the Jacobian is

$$J = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2.$$

Thus, all the four conditions of the Theorem 1.2 hold, and hence, $\boldsymbol{Y} = \left(\frac{X_1}{X_2},\,X_2\right)$ is a continuous random vector with JPDF

$$f_{Y_1,Y_2}(y_1,\,y_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(1+y_1^2)y_2^2}\,|y_2| \quad \text{for } (y_1,\,y_2) \in \mathbb{R}^2.$$

Now, we can find the marginal PDF of Y_1 from the JPDF of $(Y_1,\,Y_2)$. The marginal PDF of Y_1 is given by

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} \frac{|y_2|}{2\pi} e^{-\frac{1}{2}(1+y_1^2)y_2^2} dy_2 = \frac{1}{\pi}\int_0^{\infty} y_2 e^{-\frac{1}{2}(1+y_1^2)y_2^2} dy_2 = \frac{1}{\pi(1+y_1^2)}$$

for all $y_1 \in \mathbb{R}$. Thus, $Y_1 \sim Cauchy(0,\,1)$. ||

1.1.3 Technique 3

The Technique 3 depends on the moment generating function (MGF). Hence, first we need to define the MGF of a random vector.

Definition 1.1 (Moment Generating Function)**.** *Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a random vector. The MGF of $\boldsymbol{X}$ at $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$ is defined by*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = E\left(\exp\left(\sum_{i=1}^{n} t_i X_i\right)\right)$$

provided the expectation exists in a neighborhood of origin $\mathbf{0} = (0,\,0,\,\ldots,\,0)$.

Definition 1.2. *Two n-dimensional random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to have the same distribution, denoted by $\boldsymbol{X} \overset{d}{=} \boldsymbol{Y}$, if $F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{\boldsymbol{Y}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

Theorem 1.3. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two n-dimensional random vectors. Let $M_{\boldsymbol{X}}(\boldsymbol{t}) = M_{\boldsymbol{Y}}(\boldsymbol{t})$ for all $\boldsymbol{t}$ in a neighborhood around $\mathbf{0}$, then $\boldsymbol{X} \overset{d}{=} \boldsymbol{Y}$.*

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

Example 1.15. Let $X \sim N(\mu, \sigma^2)$. Suppose we are interested to find the distribution of $Y = a + bX$, which is a linear combination of X. Assume that $b \neq 0$. Otherwise $Y = a$ with probability one. First, let us try to find the MGF of Y. Note that

$$E\left(e^{tY}\right) = E\left(e^{t(a+bX)}\right) = e^{ta}E\left(e^{tbX}\right) = e^{ta}M_X(tb) \quad \text{for all } t \in \mathbb{R}.$$

Hence,

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

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

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

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

Example 1.16. Let X_i, $i = 1, 2, \ldots, k$ be independent $Bin(n_i, p)$ RVs. Let us try to find the distribution of $Y = \sum_{i=1}^{k} X_i$. Now, the MGF of Y is

$$M_Y(t) = E\left(e^{tY}\right) = E\left(\exp\left(t\sum_{i=1}^{k} X_i\right)\right) = E\left(\prod_{i=1}^{k} e^{tX_i}\right) = \prod_{i=1}^{k} E(e^{tX_i}) = \prod_{i=1}^{k} M_{X_i}(t).$$

The fourth equality is true as the RVs X_1, X_2, $\ldots$, X_k are independent. Note that the MGF of $X \sim Bin(n, p)$ is $M_X(t) = (1 - p + pe^t)^n$ for all $t \in \mathbb{R}$. Thus, the MGF of Y is

$$M_Y(t) = \prod_{i=1}^{k}(1 - p + pe^t)^{n_i} = (1 - p + pe^t)^{\sum_{i=1}^{k} n_i}$$

for $t \in \mathbb{R}$. Let $Z \sim Bin\left(\sum_{i=1}^{k} n_i, p\right)$, then $M_Z(t) = M_Y(t)$ for all $t \in \mathbb{R}$. Thus, $Y \stackrel{d}{=} Z \sim Bin\left(\sum_{i=1}^{k} n_i, p\right)$. Note that this example is an extension of Example 1.9. ||

Example 1.17. Let $X_1, X_2, \ldots, X_k \stackrel{i.i.d.}{\sim} Exp(\lambda)$ and $Y = \sum_{i=1}^{k} X_i$. Then the MGF of Y is

$$M_Y(t) = \prod_{i=1}^{k} M_{X_i}(t) = [M_{X_1}(t)]^k = \left(1 - \frac{t}{\lambda}\right)^{-k}$$

for all $t < \lambda$. The second equality is due to the fact that X_i has same distribution for all $i = 1, 2, \ldots, k$. Let $Z \sim Gamma(k, \lambda)$. Then $M_Z(t) = M_Y(t)$ for all $t < \lambda$. Hence, $Y \sim Gamma(k, \lambda)$. ||

Example 1.18. Let X_i, $i = 1, 2, \ldots, k$ be independent $N(\mu_i, \sigma_i^2)$ RVs. Then $\sum_{i=1}^{k} X_i \sim N\left(\sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \sigma_i^2\right)$. This can be proved following the same technique as the last example. I am leaving it as an exercise. ||

1.2 Bivariate Normal Distribution

Definition 1.3 (Expectation of a Random Vector)**.** *Expectation of a random vector is given by*

$$E(\boldsymbol{X}) = (EX_1, EX_2, \ldots, EX_n)' = \boldsymbol{\mu}.$$

Definition 1.4 (Variance-Covariance Matrix of a Random Vector)**.** *The variance-covariance matrix of a n-dimensional random vector, denoted by Σ, is defined by*

$$\Sigma = [Cov(X_i, X_j)]_{i,j=1}^{n} = E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'.$$

Definition 1.5 (Univariate Normal Distibution)**.** *A CRV X is said to have a univariate normal distribution if the PDF of X is given by*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for all } x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. In this case, $X \sim N(\mu, \sigma^2)$ is used to denote the RV X follows a normal distribution with parameters μ and σ^2.

Remark 1.2. Note that if $X \sim N(\mu, \sigma^2)$, then all moments of X exist. In particular, $E(X)$ and $Var(X)$ exist, and they are given by $E(X) = \mu$ and $Var(X) = \sigma^2$. This means that a normal distribution is completely specified by its mean and variance. †

Theorem 1.4 (MGF of Uni-variate Normal Distribution)**.** *If $X \sim N(\mu, \sigma^2)$, then the MGF of X is $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ for all $t \in \mathbb{R}$.*

Proof: The proof is straight forward from the definition of MGF. □

Definition 1.6 (Bivariate Normal)**.** *A two dimensional random vector $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is said to have a bivariate normal distribution if $aX_1 + bX_2$ is a univariate normal for all $(a, b) \in \mathbb{R}^2 \setminus (0, 0)$.*

Theorem 1.5. *If $\boldsymbol{X}$ has bivariate normal distribution, then each of X_1 and X_2 is univariate normal. Hence, $E(X_1)$, $E(X_2)$, $Var(X_1)$, $Var(X_2)$, and $Cov(X_1, X_2)$ exist.*

Proof: Taking $a = 1$ and $b = 0$, $aX_1 + bX_2 = X_1$ follows normal distribution. Similarly, X_2 follows normal distribution. As all moments of a normal RV exist, $E(X_1)$, $E(X_2)$, $Var(X_1)$, and $Var(X_2)$ exist. As $|Cov(X_1, X_2)| \leq \sqrt{Var(X_1)Var(X_2)}$, $Cov(X_1, X_2)$ exists. □

Let us denote $\boldsymbol{\mu} = E(\boldsymbol{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = Var(\boldsymbol{X}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$, where $\mu_1 = E(X_1)$, $\mu_2 = E(X_2)$, $\sigma_{11} = Var(X_1)$, $\sigma_{22} = Var(X_2)$, and $\sigma_{12} = \sigma_{21} = Cov(X_1, X_2)$.

Theorem 1.6. *Let $\boldsymbol{X}$ be a bivariate normal random vector. If $\boldsymbol{\mu} = E(\boldsymbol{X})$ and $\Sigma = Var(\boldsymbol{X})$, then for any fixed $\boldsymbol{u} = (a, b) \in \mathbb{R}^2 \setminus (0, 0)$,*

$$\boldsymbol{u}'\boldsymbol{X} \sim N(\boldsymbol{u}'\boldsymbol{\mu}, \boldsymbol{u}'\Sigma\boldsymbol{u}).$$

Proof: As $\boldsymbol{u}'\boldsymbol{X} = aX_1 + bX_2$, $\boldsymbol{u}'\boldsymbol{X}$ follows a univariate normal distribution. Now,

$$E(\boldsymbol{u}'\boldsymbol{X}) = a\mu_1 + b\mu_2 = \boldsymbol{u}'\boldsymbol{\mu}.$$

and

$$Var(\boldsymbol{u}'\boldsymbol{X}) = a^2\sigma_{11} + b^2\sigma_{22} + 2ab\sigma_{12} = \boldsymbol{u}'\Sigma\boldsymbol{u}.$$

Thus, $\boldsymbol{u}'\boldsymbol{X} \sim N(\boldsymbol{u}'\boldsymbol{\mu}, \boldsymbol{u}'\Sigma\boldsymbol{u})$. □

Theorem 1.7 (MGF of Bivariate Normal Distribution)**.** *Let $\boldsymbol{X}$ be a bivariate normal random vector with $\boldsymbol{\mu} = E(\boldsymbol{X})$ and $\Sigma = Var(\boldsymbol{X})$, then the MGF of $\boldsymbol{X}$ is given by*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = e^{\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}}$$

for all $\boldsymbol{t} \in \mathbb{R}^2$.

Proof: The JMGF of $\boldsymbol{X}$ is

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = E\left(e^{\boldsymbol{t}'\boldsymbol{X}}\right) = M_{\boldsymbol{t}'\boldsymbol{X}}(1). \tag{1.3}$$

As $\boldsymbol{X}$ has a bivariate normal distribution, $\boldsymbol{t}'\boldsymbol{X} \sim N(\boldsymbol{t}'\boldsymbol{\mu}, \boldsymbol{t}'\Sigma\boldsymbol{t})$. Now, using Theorem 1.4, the proof is immediate. □

The Theorem 1.7 shows that the bivariate normal distribution is completely specified by the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix Σ. We will use the notation $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ to denote that the random vector $\boldsymbol{X}$ follows a bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix Σ.

Theorem 1.8 (Marginal Distribution)**.** *If $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$, then $X_1 \sim N(\mu_1, \sigma_{11})$ and $X_2 \sim N(\mu_2, \sigma_{22})$.*

Proof: The proof of the theorem is immediate from Theorem 1.6. □

The converse of the Theorem 1.8 is not true in general. Consider the following example in this regard.

Example 1.19. Let $X \sim N(0, 1)$. Let Z be a DRV, which is independent of X and

$$P(Z = 1) = 0.5 = P(Z = -1).$$

Then $Y = ZX \sim N(0, 1)$. To see it, notice that for all $y \in \mathbb{R}$,

$$\begin{aligned} P(Y \leq y) &= P(ZX \leq y) \\ &= P(ZX \leq y | Z = 1)\, P(Z = 1) + P(ZX \leq y | Z = -1)\, P(Z = -1) \\ &= \frac{1}{2}P(X \leq y) + \frac{1}{2}P(X \geq -y) \\ &= \Phi(y). \end{aligned}$$

Thus, $X \sim N(0, 1)$ and $Y \sim N(0, 1)$. However, (X, Y) is not a bivariate normal random vector. To see it, observe that

$$P(X + Y = 0) = P(X + ZX = 0) = P(Z = -1) = \frac{1}{2}.$$

That means that $X + Y$ does not follow a univariate normal distribution, and hence, (X, Y) is not a bivariate normal random vector, though $X \sim N(0, 1)$ and $Y \sim N(0, 1)$. ||

Theorem 1.9. *If $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ and $Cov(X_1, X_2) = 0$, then X_1 and X_2 are independent.*

Proof: In this case, $\Sigma = diag(\sigma_{11}, \sigma_{22})$. Hence, the JMGF of (X_1, X_2) is

$$\begin{aligned} M_{X_1, X_2}(t_1, t_2) &= e^{t_1\mu_1 + \frac{1}{2}\sigma_{11}t_1^2} \times e^{t_2\mu_2 + \frac{1}{2}\sigma_{22}t_2^2} \\ &= M_{X_1}(t_1) M_{X_2}(t_2), \end{aligned}$$

where $M_{X_i}(\cdot)$ is the MGF of X_i, $i = 1, 2$. This shows that X_1 and X_2 are independent. □

Note that two random variable can be dependent even if covariance between them zero. The bivariate normal random vector is special in this respect.

Theorem 1.10 (Probability Density Function)**.** *Let $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ be such that Σ is invertible, then, for all $\boldsymbol{x} \in \mathbb{R}^2$, $\boldsymbol{X}$ has a joint PDF given by*

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}} \end{aligned}$$

where $\sigma_1 = \sqrt{\sigma_{11}}$, $\sigma_2 = \sqrt{\sigma_{22}}$, ρ is correlation coefficient between X_1 and X_2.

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

Theorem 1.11 (Conditional Probability Density Function)**.** *Let $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ be such that Σ is invertible, then for all $x_2 \in \mathbb{R}$, the conditional PDF of X_1 given $X_2 = x_2$ is given by*

$$f_{X_1|X_2}(x_1|x_2) = \frac{1}{\sigma_{1|2}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_1 - \mu_{1|2}}{\sigma_{1|2}}\right)^2\right] \quad \text{for } x_1 \in \mathbb{R},$$

where $\mu_{1|2} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ and $\sigma_{1|2}^2 = \sigma_1^2(1-\rho^2)$. Thus, $X_1|X_2 = x_2 \sim N\left(\mu_{1|2}, \sigma_{1|2}^2\right)$.

Proof: Easy to see from the fact that

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

Of course, you need to perform some algebra. □

Corollary 1.1. *Under the condition of the Theorem 1.11, $E(X_1|X_2 = x_2) = \mu_{1|2} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ and $Var(X_1|X_2 = x_2) = \sigma_{1|2}^2 = \sigma_1^2(1-\rho^2)$ for all $x_2 \in \mathbb{R}$. Hence, the conditional variance does not depend on x_2.*

Proof: Straight forward from the previous theorem. □

1.3 Some Results on Independent and Identically Distributed Normal RVs

Theorem 1.12. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(0,1)$ random variables. Then*

$$\sum_{i=1}^{n} X_i^2 \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right) \equiv \chi_n^2.$$

Proof: The MGF of X_1^2 is given by

$$M_{X_1^2}(t) = E\left(e^{tX_1^2}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(\frac{1}{2}-t)x^2} dx = (1-2t)^{-\frac{1}{2}},$$

for $t < \frac{1}{2}$. Hence, the MGF of $T = \sum_{i=1}^{n} X_i^2$

$$M_T(t) = \prod_{i=1}^{n} M_{X_i^2}(t) = (1-2t)^{-\frac{n}{2}},$$

where $t < \frac{1}{2}$. Thus, $T = \sum_{i=1}^{n} X_i^2 \sim Gamma(\frac{n}{2}, \frac{1}{2})$. This distribution is also known as χ^2 distribution with degrees of freedom n. Thus, the sum of squares of n *i.i.d.* $N(0, 1)$ RVs has a χ^2 distribution with degrees of freedom n. □

Theorem 1.13. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Then $\overline{X}$ and S^2 are independently distributed and*

$$\overline{X} \sim N(\mu, \sigma^2/n) \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Proof: Let A be an $n \times n$ orthogonal matrix, whose first row is

$$\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right).$$

Note that such a matrix exists as we can start with the row and construct a basis of $\mathbb{R}^n$. Then Gram-Schmidt orthogonalization will give us the required matrix. As A is orthogonal, its' inverse exists and $A^{-1} = A^T$, the transpose of A. Now, consider the transformation of random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$ given by

$$\boldsymbol{Y} = A\boldsymbol{X}.$$

First, we shall try to find the distribution of $\boldsymbol{Y}$. Note that the transformation $g(\boldsymbol{x}) = A\boldsymbol{x}$ is a one-to-one transformation as A is invertible. The inverse transformation is given by $\boldsymbol{x} = A'\boldsymbol{y}$. Hence, the Jacobian of the inverse transformation is $J = det(A)$. As A is orthogonal, absolute value of $det(A)$ is one. Now, as $X_1, X_2, \ldots, X_n$ are *i.i.d.* $N(\mu, \sigma^2)$ RVs, the JPDF of $\boldsymbol{X}$, for $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)' \in \mathbb{R}^n$, is

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{x}-\boldsymbol{\mu})\right], \end{aligned}$$

where $\boldsymbol{\mu} = (\mu, \mu, \ldots, \mu)'$ is a n component vector. Thus, the JPDF of $\boldsymbol{Y}$, for $\boldsymbol{y} \in \mathbb{R}^n$, is

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{\boldsymbol{X}}(A'\boldsymbol{y}) \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}(A'\boldsymbol{y} - \boldsymbol{\mu})'(A'\boldsymbol{y} - \boldsymbol{\mu})\right] \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\eta})'(\boldsymbol{y} - \boldsymbol{\eta})\right], \end{aligned}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)' = A\boldsymbol{\mu}$. Note that $\eta_1 = \sqrt{n}\mu$. Moreover,

$$\boldsymbol{\eta}'\boldsymbol{\eta} = \boldsymbol{\mu}'\boldsymbol{\mu} \implies \sum_{i=1}^{n} \eta_i^2 = n\mu^2 \implies \sum_{i=2}^{n} \eta_i^2 = n\mu^2 - \eta_1^2 = 0.$$

Thus, $\eta_i = 0$ for $i = 2, 3, \ldots, n$. Hence, the JPDF of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_1 - \sqrt{n}\mu)^2} \left\{ \prod_{i=2}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y_i^2}{2\sigma^2}} \right\} \quad \text{for } \boldsymbol{y} = (y_1, y_2, \ldots, y_n)' \in \mathbb{R}^n.$$

Therefore, $Y_1, Y_2, \ldots, Y_n$ are independent RVs and $Y_1 \sim N(\sqrt{n}\mu, \sigma^2)$ and $Y_i \sim N(0, \sigma^2)$ for $i = 2, 3, \ldots, n$, where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$. Now,

$$Y_1 = \sqrt{n}\,\overline{X} \implies \sqrt{n}\,\overline{X} \sim N(\sqrt{n}\mu, \sigma^2) \implies \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Again,

$$\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X} \implies \sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2 - Y_1^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = (n-1)S^2.$$

For $i = 2, 3, \ldots, n$, $\frac{Y_i}{\sigma}$ are *i.i.d.* $N(0, 1)$ RVs. Thus, using the previous theorem

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=2}^{n} \left(\frac{Y_i}{\sigma}\right)^2 \sim \chi^2_{n-1}.$$

Notice that $\overline{X}$ is a function of Y_1 only, and S^2 is a function of $Y_2, Y_3, \ldots, Y_n$. As Y_i's are independent, $\overline{X}$ and S^2 are independent. □

Definition 1.7 (t-distribution)**.** *A CRV X is said to have a Student's t-distribution (or simply, t-distribution) with n degrees of freedom if the PDF of X is given by*

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad \text{for } t \in \mathbb{R}.$$

We will use the notation $X \sim t_n$ to denote that the RV X has a t-distribution with n degrees of freedom.

Theorem 1.14. *Let $X \sim N(0, 1)$ and $Y \sim \chi^2_n$ be two independent RVs. Then the RV $T = \dfrac{X}{\sqrt{Y/n}} \sim t_n$.*

Proof: This theorem can be proved using the transformation technique 2. Note that the JPDF of X and Y is

$$f_{X,Y}(x,\, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \times \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} \quad \text{for } x \in \mathbb{R},\, y > 0.$$

Take $V = \sqrt{\frac{Y}{n}}$. Then the inverse mapping is $x = tv$ and $y = nv^2$. The Jacobian of the transformation is

$$J = \begin{vmatrix} v & t \\ 0 & 2nv \end{vmatrix} = 2nv^2 > 0.$$

Thus, the JPDF of T and V is

$$f_{T,V}(t,\, v) = \frac{n^{\frac{n}{2}}}{2^{\frac{n-1}{2}}\sqrt{\pi}\Gamma(\frac{n}{2})} v^n e^{-\frac{1}{2}nv^2\left(1+\frac{t^2}{n}\right)} \quad \text{for } t \in \mathbb{R},\, v > 0.$$

Therefore, for $t \in \mathbb{R}$, the marginal PDF of T is

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{T,V}(t,\, v)dv \\ &= \frac{n^{\frac{n}{2}}}{2^{\frac{n-1}{2}}\sqrt{n}\Gamma(\frac{n}{2})} \int_0^\infty v^n e^{-\frac{1}{2}nv^2\left(1+\frac{t^2}{n}\right)} dv \\ &= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}. \end{aligned}$$

□

Corollary 1.2. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Then*

$$\sqrt{n}\,\frac{\overline{X}-\mu}{S} \sim t_{n-1},$$

where S is the positive square root of S^2.

Proof: From Theorem 1.13, it is clear that $\sqrt{n}\,\frac{\overline{X}-\mu}{\sigma} \sim N(0,\, 1)$. Therefore,

$$\frac{\sqrt{n}\,\frac{\overline{X}-\mu}{\sigma}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \sqrt{n}\,\frac{\overline{X}-\mu}{S} \sim t_{n-1}.$$

□

Definition 1.8 (F-distribution)**.** *A CRV X is said to have a F-distribution with n and m degrees of freedom if the PDF of X is given by*

$$f(x) = \frac{1}{B\left(\frac{n}{2},\, \frac{m}{2}\right)} \left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1} \left(1+\frac{n}{m}x\right)^{-\frac{n+m}{2}} \quad \textit{for } x > 0.$$

We will use the notation $X \sim F_{n,m}$ to denote that the RV X has a F-distribution with n and m degrees of freedom.

Theorem 1.15. *Let $X \sim \chi^2_n$ and $Y \sim \chi^2_m$ are two independent RVs. Then*

$$F = \frac{X/n}{Y/m} = \frac{mX}{nY} \sim F_{n,\,m}.$$

Proof: The JPDF of X and Y is

$$f_{X,\,Y}(x,\,y) = \frac{1}{2^{\frac{m+n}{2}}\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} x^{\frac{n}{2}-1} y^{\frac{m}{2}-1} e^{-\frac{1}{2}(x+y)} \quad \text{for } x > 0,\, y > 0.$$

Taking $V = Y$, the inverse transformation is $x = \frac{n}{m} f v$ and $y = v$. The Jacobian of the inverse transformation is

$$J = \begin{vmatrix} \frac{n}{m} v & \frac{n}{m} f \\ 0 & 1 \end{vmatrix} = \frac{n}{m} v > 0.$$

Thus, the JPDF of F and V is

$$f_{F,\,V}(f,\,v) = \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{2^{\frac{m+n}{2}}\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} f^{\frac{n}{2}-1} v^{\frac{m+n}{2}-1} e^{-\frac{1}{2}(1+\frac{n}{m}f)v} \quad \text{for } f > 0,\, v > 0.$$

Therefore, for $f > 0$, the marginal PDF of F is

$$\begin{aligned} f_F(f) &= \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{2^{\frac{m+n}{2}}\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} f^{\frac{n}{2}-1} \int_0^\infty v^{\frac{m+n}{2}-1} e^{-\frac{1}{2}(1+\frac{n}{m}f)v} dv \\ &= \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{B\left(\frac{n}{2},\,\frac{m}{2}\right)} f^{\frac{n}{2}-1} \left(1 + \frac{n}{m} f\right)^{\frac{n+m}{2}}. \end{aligned}$$

□

Corollary 1.3. *Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2)$. Also, assume that X_i's and Y_j's are independent. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$, $\overline{Y} = \frac{1}{m}\sum_{i=1}^m Y_i$, and $S_Y^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \overline{Y})^2$. Then*

$$\frac{\sigma_2^2 S_X^2}{\sigma_1^2 S_Y^2} \sim F_{n-1,m-1}.$$

Proof: The proof is straight forward from the Theorems 1.15 and 1.13. □

1.4 Modes of Convergence

This section will deal with convergence properties of a sequence of RVs. There are several modes of convergence of sequence of RVs. Here, we will discuss four modes of convergence for a sequence of RVs $\{X_n\}$. These are quite useful concepts in probability. They have applications in different other fields including Statistics.

Definition 1.9 (Almost Sure Convergence)**.** *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. We say that X_n converges almost surely or with probability (w.p.) 1 to a random variable X if*

$$P\left(\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\}\right) = 1.$$

Example 1.20. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability (for any interval $I \subseteq \mathcal{S}$, $P(I) =$ length of I). Define the sequence of RVs by

$$X_n(\omega) = 1_{[0,\frac{1}{n}]}(\omega) \quad \text{for all } n = 1,\, 2,\, 3,\, \ldots.$$

Then X_n converges almost surely to the zero RV. Here, the zero RV means a RV, say X, defined on the same probability space $(\mathcal{S},\, \mathcal{F},\, P)$ such that $X(\omega) = 0$ for all $\omega \in \mathcal{S}$. To see it, notice that for any fixed $\omega \in (0,\, 1]$, we can find an n_0 such that $\frac{1}{n} < \omega$ for all $n \geq n_0$. Thus, $X_n(\omega) \to 0 = X(\omega)$ as $n \to \infty$. Therefore, $\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\} = (0,\, 1]$ and hence,

$$P\left(\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\}\right) = P\left((0,\, 1]\right) = 1.$$

Thus, $X_n \to 0$ almost surely. ||

Definition 1.10 (Convergence in Probability)**.** *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. We say that X_n converges in probability to a random variable X if for any $\epsilon > 0$,*

$$P(|X_n - X| > \epsilon) \to 0 \quad \textit{as } n \to \infty.$$

Example 1.21. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability. Define the sequence of RVs using $X_n = 1_{[0,\frac{1}{n}]}$. Then X_n converges in probability to the zero random variable. Let X denote the zero RV defined on the same probability space. To see it, notice that for any fixed $\epsilon > 0$, $|X_n - X| > \epsilon$ only on the interval $[0,\, \frac{1}{n}]$. Thus,

$$P\left(|X_n - X| > \epsilon\right) = \frac{1}{n} \implies \lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0.$$

Therefore, $X_n \to X$ in probability. ||

Example 1.22. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability. Define the sequence of RVs using $X_n = n1_{[0,\frac{1}{n}]}$. Then X_n converges in probability to the zero random variable. Let X denote the zero RV defined on the same probability space. To see it, notice that for any fixed $\epsilon > 0$, $|X_n - X| > \epsilon$ only on the interval $[0,\, \frac{1}{n}]$. Thus,

$$P\left(|X_n - X| > \epsilon\right) = \frac{1}{n} \implies \lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0.$$

Therefore, $X_n \to X$ in probability. ||

It may seem that convergence almost surely and convergence in probability are equivalent. However, this is not true, as the following example shows.

Example 1.23. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability. Define the sequence of RVs by

$$X_{m,n} = 1_{[\frac{m-1}{2^n},\, \frac{m}{2^n}]} \quad \text{for } m = 1,\, 2,\, \ldots,\, 2^n;\; n = 1,\, 2,\, 3,\, \ldots.$$

Note that $X_{1,1} = 1_{[0,1/2]}, X_{2,1} = 1_{[1/2,1]}, X_{1,2} = 1_{[0,1/4]}, X_{2,2} = 1_{[1/4,1/2]}, X_{3,2} = 1_{[1/2,3/4]}, X_{4,2} = 1_{[3/4,1]}$ and so on. This sequence of RVs $\{X_{m,n}\}$ can be visualized as follows (see Figure 1.3). We start with the interval $[0,\, 1]$. First, we divide the interval into two equal parts, $[0,\, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. The first RV $X_{1,1}$ is 1 on the first part and 0 on the second part. The second RV $X_{2,1}$ is 1 on the second part and 0 on the first part. Then, we divide the interval into 2^2 equal

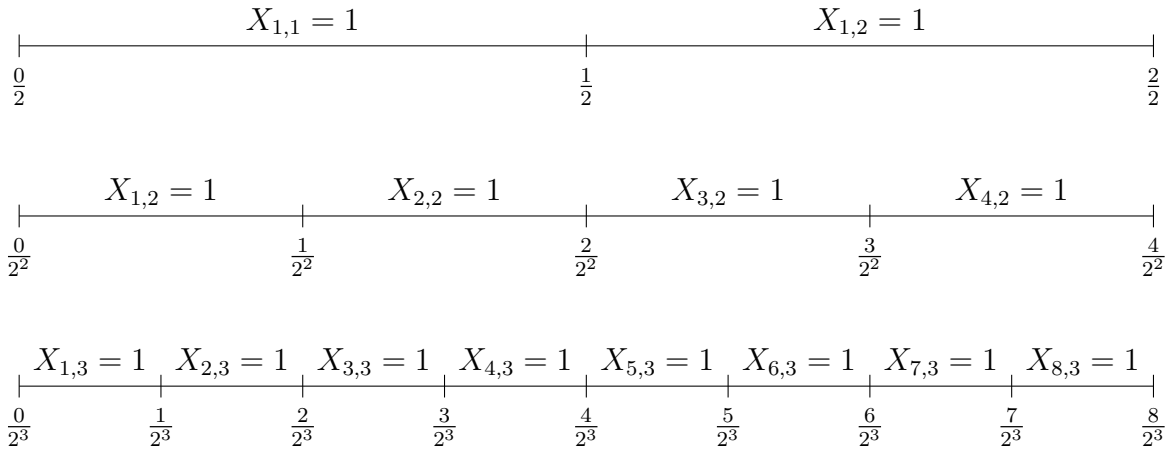


Figure 1.3: Figure for Example 1.23

parts, *viz.*, $[0, \frac{1}{2^2}]$, $[\frac{1}{2^2}, \frac{2}{2^2}]$, $[\frac{2}{2^2}, \frac{3}{2^2}]$, and $[\frac{3}{2^2}, 1]$. Now, the third RV $X_{1,2}$ is 1 on the first part $[1, \frac{1}{4}]$ and 0 otherwise. The fourth RV $X_{2,2}$ is 1 on the second part $[\frac{1}{4}, \frac{1}{2}]$ and 0 otherwise. The fifth RV $X_{3,2}$ equals 1 on the third part $[\frac{1}{2}, \frac{3}{4}]$ and 0 otherwise. Finally, the sixth RV $X_{4,2}$ is 1 on the fourth part $[\frac{3}{4}, 1]$ and 0 otherwise. Next, we divide the interval $[0, 1]$ into 2^3 equal parts and define the next 8 RVs in the similar manner. This procedure continues.

Let us assume that X be a RV defined on the same probability space and $X = 0$. Then, for any $\epsilon > 0$,

$$P\left(|X_{m,n} - X| > \epsilon\right) = \frac{1}{2^n} \implies \lim_{n\to\infty} P\left(|X_{m,n} - X| > \epsilon\right) = 0.$$

Therefore, $X_{m,n} \to X$ in probability. However, for any fixed $\omega \in \mathcal{S}$, there exists a subsequence of the sequence of real numbers $\{X_{m,n}(\omega)\}$ that converges to one and another subsequence that converges to zero. Therefore, $\{X_{m,n}(\omega)\}$ does not converge for all $\omega \in \mathcal{S}$. Thus,

$$P\left(\{\omega \in \mathcal{S} : X_{m,n} \text{ converges}\}\right) = P\left(\emptyset\right) = 0.$$

This shows that $X_{m,n}$ do not converge to any RV almost surely. This example shows that a sequence of RVs, which converges in probability, may not converge almost surely. ||

Definition 1.11 (Convergence in r^{th} Mean)**.** *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. For $r = 1, 2, 3, \ldots$, we say that X_n converges in r^{th} mean to a random variable X if*

$$E|X_n - X|^r \to 0 \quad \text{as } n \to \infty.$$

Example 1.24. Let $\mathcal{S} = [0, 1], \mathcal{F} = \mathcal{B}([0, 1])$, and P be a uniform measure. Define $X_n = 1_{[0,\frac{1}{n}]}$. Then X_n converges in 1st mean to the zero random variable. To see it, notice that

$$E|X_n - X| = \frac{1}{n} \to 0 \quad \text{as } n \to \infty,$$

where X is a zero RV defined on the same probability space. ||

Definition 1.12 (Convergence in Distribution)**.** *Let $\{X_n\}$ be a sequence of RVs and X be a RV. Let $F_n(\cdot)$ and $F(\cdot)$ denote the CDF of X_n and X, respectively. We say that X_n converges in distribution to a random variable X if*

$$F_n(x) \to F(x) \quad \text{as } n \to \infty$$

for all x where F is continuous.

Unlike the first three modes of convergence, here X_n's can be defined on different probability spaces. We are only interested if the sequence of CDFs converges to a CDF. This flexibility makes this mode of convergence very useful.

Example 1.25. Suppose X_n's are random variables such that $P(X_n = \frac{1}{n}) = 1$. Then, the CDF of X_n is

$$F_n(x) = \begin{cases} 0 & \text{if } x < \frac{1}{n} \\ 1 & \text{if } x \geq \frac{1}{n}, \end{cases}$$

which converges pointwise to the function

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

for all $x \neq 0$, which is the point of discontinuity of the function $F(\cdot)$. Now, $F(\cdot)$ is the CDF of the RV X, which takes value 0 with probability one. Therefore, X_n converges in distribution to the zero RV. ||

The following theorems states the relation between different modes of convergence.

Theorem 1.16. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. Then $X_n \to X$ in probability if $X_n \to X$ almost surely.*

Proof: This prove is skipped here. □

Theorem 1.17. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. Then $X_n \to X$ in probability if $X_n \to X$ in rth mean for any $r = 1, 2, 3, \ldots$.*

Proof: Let $X_n \to X$ in rth mean. Then, using Markov inequality, for any $\epsilon > 0$,

$$P\left(|X_n - X| > \epsilon\right) \leq \frac{E|X_n - X|^r}{\epsilon^r} \to 0 \quad \text{as } n \to \infty.$$

As probability of an event is always non-negative,

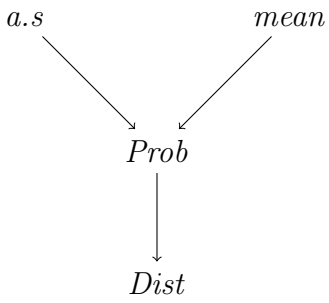
$$P\left(|X_n - X| > \epsilon\right) \to 0 \quad \text{as } n \to \infty.$$

Thus, $X_n \to X$ in probability. □

Theorem 1.18. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. Then $X_n \to X$ in distribution if $X_n \to X$ in probability.*

Proof: The proof is skipped here. □

The following figure depicts the relationship between several modes of convergence pictorially. Note that the arrows are one-sided. What about other sides? Moreover, there is no arrows between almost sure convergence and rth mean convergence. The following examples show that in general one mode of convergence does not imply other, whenever there is no directed arrows in the above figure. The Example 1.23 shows that probability convergence does not imply almost sure convergence.



Example 1.26. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability. Define the sequence of RVs by

$$X_{m,n} = 1_{[\frac{m-1}{2^n}, \frac{m}{2^n}]} \quad \text{for } m = 1, 2, \ldots, 2^n; n = 1, 2, 3, \ldots.$$

Then

$$E|X_{m,n}| = \frac{1}{2^n} \to 0 \quad \text{as } n \to \infty.$$

Thus, $X_{m,n} \to X = 0$ in 1st mean. However, in Example 1.23, we have seen that $X_{m,n}$ does not convergence almost surely. This example shows that rth mean convergence does not imply almost sure convergence. ||

Example 1.27. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$ and P be a uniform probability. Define $X_n = n1_{[0,\frac{1}{n}]}$. Now, taking $X = 0$,

$$P\left(|X_n - X| > \epsilon\right) = \frac{1}{n} \to 0 \quad \text{as } n \to \infty,$$

for any $\epsilon > 0$. Thus, $X_n \to X$ in probability. Using the logic used in Example 1.20,

$$P\left(\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\}\right) = P((0,\,1]) = 1.$$

Thus, $X_n \to X$ almost surely. However, X_n does not converge to X in rth mean. To see it, notice that

$$E|X_n - X|^r = n^{r-1} \to \begin{cases} 1 & \text{if } r = 1 \\ \infty & \text{if } r > 1. \end{cases}$$

This example shows that probability convergence or almost sure convergence do not imply rth mean convergence. ||

Example 1.28. Let X be a $N(0,1)$ RV defined on some probability space $(\mathcal{S}, \mathcal{F}, P)$. Define $X_n = X$ for all n. Notice that the CDFs of X_n are same for all $n = 1, 2, \ldots$ and is given by $\Phi(\cdot)$. Moreover, the CDFs of X and $-X$ are also $\Phi(\cdot)$. Thus, X_n converges in distribution to $-X$. However, X_n does not converge to $-X$ in probability. To see it, we can proceed as follows: for $\epsilon > 0$,

$$P\left(|X_n + X| \leq \epsilon\right) = P\left(2|X| \leq \epsilon\right) = 2\Phi\left(\frac{\epsilon}{2}\right) - 1 \neq 1.$$

This example shows that distribution convergence does not imply probability convergence, even if the random variables are defined on the same probability space. ||

Theorem 1.19. *Suppose $\{X_n\}$ is a sequence of RVs defined on a probability space and X_n converges in distribution to some constant c, then X_n also converges in probability to c.*

Proof: As X_n converges to a constant c,

$$F_n(x) \to F(x) = \begin{cases} 0 & \text{if } x < c \\ 1 & \text{if } x \geq c \end{cases}$$

as $n \to \infty$. Now, fix $\varepsilon > 0$. Then,

$$\begin{aligned} 0 \leq P\left(|X_n - c| > \varepsilon\right) &= P\left(X_n > c + \varepsilon\right) + P\left(X_n < c - \varepsilon\right) \\ &\leq 1 - F_n(c + \varepsilon) + F_n(c - \varepsilon) \to 1 - 1 + 0 = 0 \end{aligned}$$

as $n \to \infty$. Note that as $c + \varepsilon > c$ and $c - \varepsilon < c$, $F_n(c + \varepsilon) \to 1$ and $F_n(c - \varepsilon) \to 0$. Thus, $X_n \to c$ in probability. □

Corollary 1.4. *Suppose $\{X_n\}$ is a sequence of RVs defined on a probability space. Then, $X_n \to c$ in distribution if and only if $X_n \to c$ in probability, where c is a constant.*

Proof: The proof of the corollary is straight forward by combining the previous theorem and Theorem 1.18. □

The following theorems provide several properties of different modes of convergence. The proof of the theorems are skipped here.

Theorem 1.20. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Suppose $X_n \to X$ w. p. 1 and $Y_n \to Y$ w. p. 1. Then*

- *$X_n + Y_n \to X + Y$ w. p. 1.*
- *$X_n Y_n \to XY$ w. p. 1.*
- *$f(X_n) \to f(X)$ w. p. 1, for any f continuous.*

Theorem 1.21. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Suppose $X_n \to X$ in probability and $Y_n \to Y$ in probability. Then*

- *$X_n + Y_n \to X + Y$ in probability.*
- *$X_n Y_n \to XY$ in probability.*
- *$f(X_n) \to f(X)$ in probability, for any f continuous.*

Theorem 1.22. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$.*

- *If $X_n \to X$ in r^{th} mean and $Y_n \to Y$ in r^{th} mean, then $X_n + Y_n \to X + Y$ in r^{th} mean.*
- *If $X_n \to X$ in r^{th} mean then $f(X_n) \to f(X)$ in r^{th} mean, for any f bounded continuous.*

Theorem 1.23. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Suppose $X_n \to X$ in distribution and $Y_n \to c$ in probability for some constant c. Then*

- $X_n + Y_n \to X + c$ *in distribution.*
- $X_n Y_n \to cX$ *in distribution.*
- $f(X_n) \to f(X)$ *in distribution, for any* f *continuous.*

Example 1.29. Let $X, Y \sim N(0, 1)$ and X and Y be independent RVs. Take $X_n = X$ and $Y_n = Y$ for all $n = 1, 2, 3, \ldots$. Then, $X_n \to X$ in distribution and $Y_n \to X$ in distribution. Now, $X_n + Y_n = X + Y \sim N(0, 2)$ and $2X \sim N(0, 4)$. Thus, $X_n + Y_n$ does not converges to $2X$ in distribution. This example shows that $X_n + Y_n$ may not converge to $X + Y$ in distribution if $X_n \to X$ in distribution and $Y_n \to Y$ in distribution. You can easily check that the same conclusion is also true for product. ||

Theorem 1.24. *Let* X_n *be a RV with MGF* $M_n(t)$ *for* $n = 1, 2, 3, \ldots$. *Let* X *be a RV with MGF* $M(t)$. *If* $M_n(t) \to M(t)$ *for all* t *in an open interval containing zero, as* $n \to \infty$, *then* $X_n \to X$ *in distribution.*

Theorem 1.25. *Let* X_n *be a DRV with PMF* $f_n(\cdot)$ *for* $n = 1, 2, 3, \ldots$. *Let* X *be a DRV with PMF* $f(\cdot)$. *If, for all* $x \in \mathbb{R}$, $f_n(x) \to f(x)$ *as* $n \to \infty$, *then* $X_n \to X$ *in distribution.*

Theorem 1.26. *Let* X_n *be a CRV with PDF* $f_n(\cdot)$ *for* $n = 1, 2, 3, \ldots$. *Let* X *be a CRV with PDF* $f(\cdot)$. *If, for all* $x \in \mathbb{R}$, $f_n(x) \to f(x)$ *as* $n \to \infty$, *then* $X_n \to X$ *in distribution.*

Example 1.30. Let $X_n \sim Bin(n, p_n)$, where $p_n \to 0$ and $np_n = \lambda\,(> 0)$. Then, for $n = 1, 2, 3, \ldots$, the MGF of X_n is

$$M_n(t) = \left(1 - p_n + p_n e^t\right)^n = \left(1 + \frac{\lambda}{n}\left(e^t - 1\right)\right)^n \to e^{\lambda\left(e^t - 1\right)}$$

for all $t \in \mathbb{R}$. Note that if $X \sim Poi(\lambda)$, then the MGF of X is

$$M(t) = e^{\lambda\left(e^t - 1\right)} \quad \text{for } t \in \mathbb{R}.$$

Thus, $X_n \to X$ in distribution.

This example tells us the motivation behind the Poisson distribution. We can use Poisson distribution to approximate the probability of a Binomial distribution when probability of success is very small and number of trials is very large. ||

Example 1.31. Under the conditions of the previous example, we can prove that $X_n \to X$ in distribution using Theorem 1.25. To see it, we can proceed as follows.

$$\begin{aligned} P\left(X_n = k\right) &= \binom{n}{k} p_n^k \left(1 - p_n\right)^{n-k} \\ &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \times \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &\to \frac{e^{-\lambda}\lambda^k}{k!}. \end{aligned}$$

Notice that the support of X_n is the set $\{0, 1, 2, \ldots, n\}$. When $n \to \infty$, the support becomes $\{0, 1, 2, \ldots\}$. ||

Example 1.32. Let $X_n \sim U(0, 1 + 1/n)$ for $n = 1, 2, 3, \ldots$. Then the PDF of X_n is

$$f_n(x) = \begin{cases} \frac{1}{1+\frac{1}{n}} & \text{if } 0 < x < 1 + \frac{1}{n} \\ 0 & \text{otherwise} \end{cases} \longrightarrow f(x) = \begin{cases} 1 & \text{if } 0 < x \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

which is the PDF of a RV $X \sim U(0, 1)$. Thus, $X_n \to X$ in distribution. ||

1.5 Limit Theorems

In this section, we will discuss two very famous and useful theorems, *viz.*, strong law of large numbers (SLLN) and central limit theorem (CLT). We will skip the proofs, but we will see some applications.

Theorem 1.27 (Strong Law of Large Numbers)**.** *Let $\{X_n\}$ be a sequence of i.i.d. RVs with finite mean μ. Define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\{\overline{X}_n\}$ converges to μ almost surely.*

Proof: The proof is skipped. □

Let us loosely discuss the intuitive idea of the previous theorem. Suppose that we want to find the average height of all Indians. Ideally, we need to go to each and every Indian and record their height. Finally, the average should be calculated based on the observations on height. This average is called population average or population mean. It is a very costly (in terms of money and time) process. Alternatively, we can take a representative sample of the Indian population. Here, sample represents a subset of original population. Then, we can collect the height data for each and every person in the sample and then calculate the mean of those sample observations. This mean is called sample mean. If the number of persons in the sample is very small (say, 5 or 10), the calculated sample mean may not be close to the original population mean. However, if we keep on increasing the sample size (the number of persons in the sample), the sample mean should get closer to population mean. The above theorem provided theoretical justification of this intuitive idea. Note that μ and $\overline{X}$ are population and sample means, respectively. Thus, loosely speaking, the SLLN states that sample mean converges to population mean almost surely as we increase the sample size.

Example 1.33 (Bernoulli proportion converges to success probability)**.** Suppose that a sequence of independent trials is performed. Let E be a fixed event. Letting

$$X_i = \begin{cases} 1 & \text{if } E \text{ occurs on the } i\text{th trial} \\ 0 & \text{if } E \text{ does not occur on the } i\text{th trial,} \end{cases}$$

we have by the SLLN that, with probability one,

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu = E(X_1) = P(E).$$

Since, $X_1 + X_2 + \ldots + X_n$ represents the number of times that the event E occurs in the first n trials, we may interpret it as stating that, with probability one, the limiting proportion of time that the event E occurs is $P(E)$. ||

Example 1.34 (Monte Carlo Integration)**.** Suppose that we want to integrate

$$I = \int_a^b h(x)dx.$$

If we cannot do it explicitly, we can use numerical technique like Simpson's $\frac{1}{3}$rd rule. Here, we will see another technique based on the SLLN. Suppose that a and b are finite real numbers. Note that the above integration can be rewritten as

$$I = (b-a)\int_a^b h(x)\frac{1}{b-a}dx = (b-a)E\left(Y\right),$$

where $Y = h(X)$ and $X \sim U(a,\, b)$. Let $\{X_n\}$ be a sequence of *i.i.d.* RVs with common distribution $U(a,\, b)$ and assume that $Y_n = h(X_n)$ for $n = 1,\, 2,\, 3,\, \ldots$. Now, SLLN says that, with probability one,

$$\overline{Y}_n = \frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \to E(Y) = \frac{I}{b-a} \implies \frac{b-a}{n}\sum_{i=1}^{n} h(X_i) \to I.$$

Thus, we can generate N random numbers from $U(a,\, b)$ and then, the integration I can be approximated by $\frac{b-a}{N}\sum_{i=1}^{N} h(X_i)$. Here, N is a large integer (the popular choices are 5000 or 10000). The generation from $U(a,\, b)$ can be done using any standard software like R, MATLAB, python etc. ||

Theorem 1.28 (Central Limit Theorem)**.** *Let $\{X_n\}$ be a sequence of i.i.d. RVs with mean μ and variance $\sigma^2 < \infty$. Then, as $n \to \infty$,*

$$P\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \le a\right) \to \Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

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

Proof: The proof is skipped. □

The central limit theorem (CLT) says that

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \to Z \sim N(0,\, 1) \quad \text{in distribution.}$$

Thus, the CDF of standardized sample mean can be approximated (for large sample size) using the CDF of a standard normal distribution, whenever X_n's are i.i.d. RVs with finite mean μ and finite variance σ^2. In other words, the CDF of sample mean can be approximated using the CDF of a $N(\mu,\, \frac{\sigma^2}{n})$ distribution. Note that CLT holds true for any distribution of X_n as long as the variance is finite.

Example 1.35 (Normal Approximation to the Binomial)**.** Let $X_n \sim Bin(n,\, p)$. Then

$$P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \le a\right) \to \Phi(a) \quad \text{as } n \to \infty.$$

We will use CLT to prove this statement. Let $\{Y_n\}$ be a sequence of i.i.d. RVs where $Y_1 \sim Bernoulli(p)$. Then, we know that

$$\sum_{i=1}^{n} Y_i \stackrel{d}{=} X_n \implies \overline{Y}_n \stackrel{d}{=} \frac{X_n}{n}.$$

Now, $E(Y_n) = p$ and $Var(Y_n) = p(1-p)$ for all $n = 1, 2, 3, \ldots$. Thus,

$$P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \leq a\right) = P\left(\sqrt{n}\frac{\overline{Y}_n - p}{\sqrt{p(1-p)}} \leq a\right) \to \Phi(a) \quad \text{as } n \to \infty.$$

The equality in the above line is due to the fact that $\overline{Y}_n$ and $\frac{X_n}{n}$ have same distribution. The convergence is due to the CLT. ||

Example 1.36. The lifetimes of a special type of battery is a RV with mean 40 hours and standard deviation 20 hours. A battery is used until it fails, at which point it is replaced by a new one. Assume a stockpile of 25 such batteries, the lifetimes of which are independent, we want to approximate the probability that over 1100 hours of use can be obtained. Let X_i denote the lifetime of the ith battery to be put in use. Then, we are interested in

$$p = P\left(X_1 + X_2 + \ldots + X_{25} > 1100\right),$$

which can be approximated as follows:

$$\begin{aligned}
p &= P\left(X_1 + X_2 + \ldots + X_{25} > 1100\right) \\
&= P\left(\overline{X}_{25} > 44\right) \\
&= P\left(\sqrt{25}\,\frac{\overline{X}_{25} - 40}{20} > \sqrt{25}\,\frac{44 - 40}{20}\right) \\
&\approx P\left(Z > 1\right), \text{ where } Z \sim N(0,\, 1). \text{ This is due to CLT.} \\
&= 1 - \Phi(1) \approx 0.1587,
\end{aligned}$$

as $\Phi(1) \approx 0.8413$. This values can be found from the normal table. ||