# Chapter 5

# Brief Overview of Probability

## OBJECTIVE OF THE CHAPTER

The principles of machine learning are largely dependent on effectively handling the uncertainty in data and predicting the outcome based on data in hand. All the learning processes we will be discussing in later chapters of this book expects the readers to understand the foundational rules of probability, the concept of random and continuous variables, distribution, and sampling principles and few basic principles such as central limit theorem, hypothesis testing, and Monte Carlo approximations. As these rules, theorems, and principles form the basis of learning principles, we will be discussing those in this chapter with examples and illustrations.

## 5.1 INTRODUCTION

As we discussed in previous chapters, machine learning provides us a set of methods that can automatically detect patterns in data, and then can be used to uncover patterns to predict future data, or to perform other kinds of decision making under uncertainty. The best way to perform such activities on top of huge data set known as **big data** is to use the tools of probability theory because probability theory can be applied to any situation involving uncertainty. In machine learning there may be uncertainties in different forms like arriving at the best prediction of future given the past data, arriving at the best model based on certain data, arriving at the confidence level while predicting the future outcome based on past data, etc. The probabilistic approach used in machine learning is closely related to the field of statistics, but the emphasis is in a different direction as we see in this chapter. In this chapter we will dis-

cuss the tools, equations, and models of probability that are useful for machine learning domain.

## 5.2 IMPORTANCE OF STATISTICAL TOOLS IN MACHINE LEARNING

In machine learning, we train the system by using a limited data set called 'training data' and based on the confidence level of the training data we expect the machine learning algorithm to depict the behaviour of the larger set of actual data. If we have observation on a subset of events, called 'sample', then there will be some uncertainty in attributing the sample results to the whole set or population. So, the question was how a limited knowledge of a sample set can be used to predict the behaviour of a real set with some confidence. It was realized by mathematicians that even if some knowledge is based on a sample, if we know the amount of uncertainty related to it, then it can be used in an optimum way without causing loss of knowledge. Refer Figure 5.1


**FIG. 5.1** Knowledge and uncertainty

Probability theory provides a mathematical foundation for quantifying this uncertainty of the knowledge. As the knowledge about the training data comes in the form of **interdependent** feature sets, the conditional probability theories (especially the Bayes theorem), discussed later in this chapter form the basis for deriving required confidence level of the training data.

Different distribution principles discussed in this chapter create the view of how the data set that we will be dealing with in machine learning can behave, in terms of their feature distributions. It is important to understand the mathematical function behind each of these distributions so that we can understand how the data is spread out from its average value – denoted by the mean and the variance. While choosing the samples from these distribution sets we should be able to calculate to what

extent the sample is representing the actual behaviour of the full data set. These along with the test of hypothesis principles build the basis for finding out the uncertainty in the training data set to represent the actual data set which is the fundamental principle of machine learning.

## 5.3 CONCEPT OF PROBABILITY – FREQUENTIST AND BAYESIAN INTERPRETATION

The concept of probability is not new. In our day to day life, we use the concept of probability in many places. For example, when we talk about probabilities of getting the heads and the tails when a coin is flipped are equal, we actually intend to say that if a coin is flipped many times, the coin will land heads a same number of times as it lands tails. This is the **frequentist** interpretation of probability. This interpretation represents the long run frequencies of events. Another important interpretation of probability tries to quantify the uncertainty of some event and thus focuses on information rather than repeated trials. This is called the **Bayesian** interpretation of probability. Taking the same example of coin flipping is interpreted as – the coin is equally likely to land heads or tails when we flip the coin next.

The reason the Bayesian interpretation can be used to model the uncertainty of events is that it does not expect the long run frequencies of the events to happen. For example, if we have to compute the probability of Brazil winning 2018 football world cup final, that event can happen only once and can't be repeated over and over again to calculate its probability. But still, we should be able to quantify the uncertainty about the event and which is only possible if we interpret probability the Bayesian way. To give some more machine learning oriented examples, we are starting a new software implementation project for a large customer and want to compute the probability of this project getting into customer escalation based on data from similar projects in the past, or we want to compute the probability of a tumor to be malignant or not based on the probability distribution of such cases among the patients of similar profile. In all these cases it is not possible to do a repeated trial of the event and the

Bayesian concept is valid for computing the uncertainty. So, in this book, we will focus on the Bayesian interpretation to develop our machine learning models. The basic formulae of probability remain the same anyway irrespective of whatever interpretation we adopt.

### 5.3.1 A brief review of probability theory

We will briefly touch upon the basic probability theory in this section just as a refresher so that we can move to the building blocks of machine learning way of use of probability. As discussed in previous chapters, the basic concept of machine learning is that we want to have a limited set of 'Training' data that we use as a representative of a large set of Actual data and through probability distribution we try to find out how an event which is matching with the training data can represent the outcome with some confidence.

#### 5.3.1.1 Foundation rules

We will review the basic rules of probability in this section.

Let us introduce few notations that will be used throughout this book.

$p(A)$ denotes the probability that the event $A$ is true. For example, $A$ might be the logical statement 'Brazil is going to win the next football world cup final'. The expression $0 \leq p(A) \leq 1$ denotes that the probability of this event happening lies between 0 and 1, where $p(A) = 0$ means the event will definitely not happen, and $p(A) = 1$ means the event will definitely happen. The notation $p(\bar{A})$ denotes the probability of the event not $A$; this is defined as $p(\bar{A}) = 1 - p(A)$. It is also common practice to write $A = 1$ to mean the event $A$ is true, and $A = 0$ to mean the event $A$ is false. So, this is a binary event where the event is either true or false but can't be something indefinite.

The probability of selecting an event $A$, from a sample size of $X$ is defined as

$p(A) = \frac{n}{X}$, where n is the number of times the instance of event $A$ is present in the sample of size $X$.

#### 5.3.1.2 Probability of a union of two events

Two events A and B are called mutually exclusive if they can't happen together. For example, England winning the Football World Cup 2018 and Brazil winning the Football World Cup 2018 are two mutually exclusive events and can't happen together. For any two events, $A$ and $B$, we define the probability of $A$ or $B$ as

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \quad (5.1)$$

$$= p(A) + p(B), \text{ if Aand Bare mutually exclusive} \quad (5.2)$$

#### 5.3.1.3 Joint probabilities

The probability of the joint event $A$ and $B$ is defined as the **product rule**:

$$p(A, B) = p(A \cap B) = p(A|B).p(B) \quad (5.3)$$

where p(A|B) is defined as the conditional probability of event A happening if event B happens. Based on this **joint distribution** on two events $p(A, B)$,we can define the **marginal distribution** as follows:

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b) \quad (5.4)$$

summing up the all probable states of B gives the total probability formulae, which is also called **sum rule or the rule of total probability**.

Same way, $p(B)$ can be defined as

$$p(B) = \sum_a p(A, B) = \sum_a p(B|A = a)p(A = a) \quad (5.5)$$

This formula can be extended for the countably infinite number of events in the set and **chain rule** of probability can be derived if the product rule is applied multiple times as

$$p(X_{1:N}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1) \\ p(X_4|X_1, X_2, X_3)\ldots p(X_N|X_{1:N} - 1) \quad (5.6)$$

#### 5.3.1.4 Conditional probability

We define the **conditional probability** of event *A*, given that event *B* is true, as follows:

$$p(A|B) = p(A, B)/p(B), \text{ if} p(B) > 0 \quad (5.7)$$

where, $p(A, B)$ is the joint probability of *A* and *B* and can also be denoted as $p(A \cap B)$

Similarly,

$$p(B|A) = p(B, A)/p(A), \text{ if} p(A) > 0 \quad (5.8)$$

**Illustration.** In a toy-making shop, the automated machine produces few defective pieces. It is observed that in a lot of 1,000 toy parts, 25 are defective. If two random samples are selected for testing without replacement (meaning that the first sample is not put back to the lot and thus the second sample is selected from the lot size of 999) from the lot, calculate the probability that both the samples are defective.

**Solution**: Let A denote the probability of first part being defective and B denote the second part being defective. Here, we have to employ the conditional probability of the second part being found defective when the first part is already found defective.

By law of probability, $p(A) = \frac{25}{1000} = 0.025$

As we are selecting the second sample without replacing the first sample into the lot and the first one is already found defective, there are now 24 defective pieces out of 999 pieces left in the lot.

Thus $$p(B|A) = \frac{24}{999} = 0.024$$

As
$$\begin{aligned} p(A, B) &= p(A \cap B) = p(B|A)p(A) \\ &= 0.025 \times 0.024 \\ &= 0.006006 \end{aligned}$$

which is the probability of both the parts being found defective.

#### 5.3.1.5 Bayes rule

The **Bayes rule**, also known as **Bayes Theorem,** can be derived by combining the definition of conditional probability with the product and sum rules, as below:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{5.9}$$

$$p(A = a|B = b) = \frac{p(A = a,\ B = b)}{p(B = b)}$$

$$= \frac{p(A = a)p(B = b|A = a)}{\sum_{a} p(A = a)p(B = b|A = a)} \tag{5.10}$$

**Illustration: Flagging an email as spam based on the sender name of the email**

Let's take an example to identify the probability of an email to be really spam based on the name of the sender. We often receive email from mail id containing junk characters or words such as bulk, mass etc. which turn out to be a spam mail. So, we want our machine learning agent to flag the spam emails for us so that we can easily delete them.

Let's also assume that we have knowledge about the reliability of the assumption that emails with sender names 'mass' and 'bulk' are spam email is 80% meaning that if some email has sender name with 'mass' or 'bulk' then the email will be spam with probability 0.8. The probability of false alarm is 10% meaning that the agent will show the positive result as spam even if the email is not a spam with a probability 0.1. Also, we have a prior knowledge that only 0.4% of the total emails received are spam.

**Solution:**

Let $x$ be the event of the flag being set as spam because the sender name has the words 'mass' or 'bulk' and $y$ be the event of some mail really being spam.

So, $p(x = 1 \mid y = 1) = 0.8$, which is the probability of the flag being positive if the email is spam.

Many times this probability is wrongly interpreted as the probability of an email being spam if the flag is positive. But that is not true, because we will then ignore the prior knowledge that only 0.4% of the emails are actually spam and also there can be a false alarm in 10% of the cases.

So, $p(y = 1) = 0.004$ and $p(y = 0) = 1 - p(y = 1) = 0.996$

$$p(x = 1 \mid y = 0) = 0.1$$

Combining these terms using the Bayes rule, we can compute the correct answer as follows:

$$\begin{aligned} p(y = 1|x = 1) &= \frac{p(x = 1|y = 1)p(y = 1)}{p(x = 1|y = 1)p(y = 1) + p(x = 1|y = 0)p(y = 0)} \\ &= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} \\ &= 0.031 \end{aligned}$$

This means that if we combine our prior knowledge with the actual test results then there is only about a 3% chance of emails actually being spam if the sender name contains the terms ‘bulk’ or ‘mass’. This is significantly lower than the earlier assumption of 80% chance and thus calls for additional checks before someone starts deleting such emails based on only this flag.

*Details of the practice application of Bayes theorem in machine learning is discussed in Chapter 6.*

## 5.4 RANDOM VARIABLES

Let’s take an example of tossing a coin once. We can associate random variables $X$ and $Y$ as

$X(\text{H}) = 1$, $X(\text{T}) = 0$, which means this variable is associated with the outcome of the coin facing head.

$Y(\text{H}) = 0$, $Y(\text{T}) = 1$, which means this variable is associated with the outcome of the coin facing tails.

Here in the sample space $S$ which is the outcome related to tossing the coin once, random variables represent the single-valued real function $[X(\zeta)]$ that assigns a real number, called its value to each sample point of $S$. A random variable is not a variable but is a function. The sample space $S$ is called the domain of random variable $X$ and the collection of all the numbers, i.e. values of $X(\zeta)$, is termed the range of the random variable.

We can define the event $(X = x)$ where $x$ is a fixed real number as

$$(X = x) = \{ \zeta : X(\zeta) = x \}$$

Accordingly, we can define the following events for fixed numbers $x$, $x_1$, and $x_2$:

$$(X \leq x) = \{ \zeta : X(\zeta) \leq x \}$$
$$(X > x) = \{ \zeta : X(\zeta) > x \}$$
$$(x_1 < X \leq x_2) = \{ \zeta : x_1 < X(\zeta) \leq x_2 \}$$

The probabilities of these events are denoted by

$$P(X = x) = P\{ \zeta : X(\zeta) = x \}$$
$$P(X \leq x) = P\{ \zeta : X(\zeta) \leq x \}$$
$$P(X > x) = P\{ \zeta : X(\zeta) > x \}$$
$$P(x_1 < X \leq x_2) = P\{ \zeta : x_1 < X(\zeta) \leq x_2 \}$$

We will use the term cdf in this book to denote the **distribution function** or **cumulative distribution function,** which takes the form of random variable X as

$$F_x(x) = P(X \leq x) \quad -\infty < x < \infty$$

Some of the important properties of $F_x(x)$ are

$$0 \leq F_x(x) \leq 1$$
$$F_x(x_1) \leq F_x(x_2) \quad \text{if } x_1 < x_2$$
$$\lim_{n \to \infty} F_x(x) = F_x(\infty) = 1$$
$$\lim_{n \to \infty} F_x(x) = F_x(-\infty) = 0$$

### 5.4.1 Discrete random variables

Let us extend the concept of binary events by defining a **discrete random variable** *X*. Let *X* be a random variable with cdf $F_x(x)$ and it changes values only in jumps (a countable number of them) and remains constant between the jumps then it is called a discrete random variable. So, the definition of a discrete random variable is that its range or the set *X* contains a finite or countably infinite number of points.

Let us consider that the jumps of $F_x(x)$ of the discrete random variable $X$ is occurring at points $x_1, x_2, x_3 \ldots$, where this sequence represents a finite or countably infinite set and $x_i < x_j$ if $i < j$. See Figure 5.2.

$$\text{Then } F_x(x_i) - F_x(x_{i-1}) = P(X \leq x_i) - P(X \leq x_{i-1}) = P(X = x_i) \qquad (5.11)$$

We can denote this through notation as

$$p(x) = p(X = x) \qquad (5.12)$$

The probability of the event that $X = x$ is denoted as $p(X = x)$, or just $p(x)$ for short. $p$ is called a **probability mass function (pmf)**.

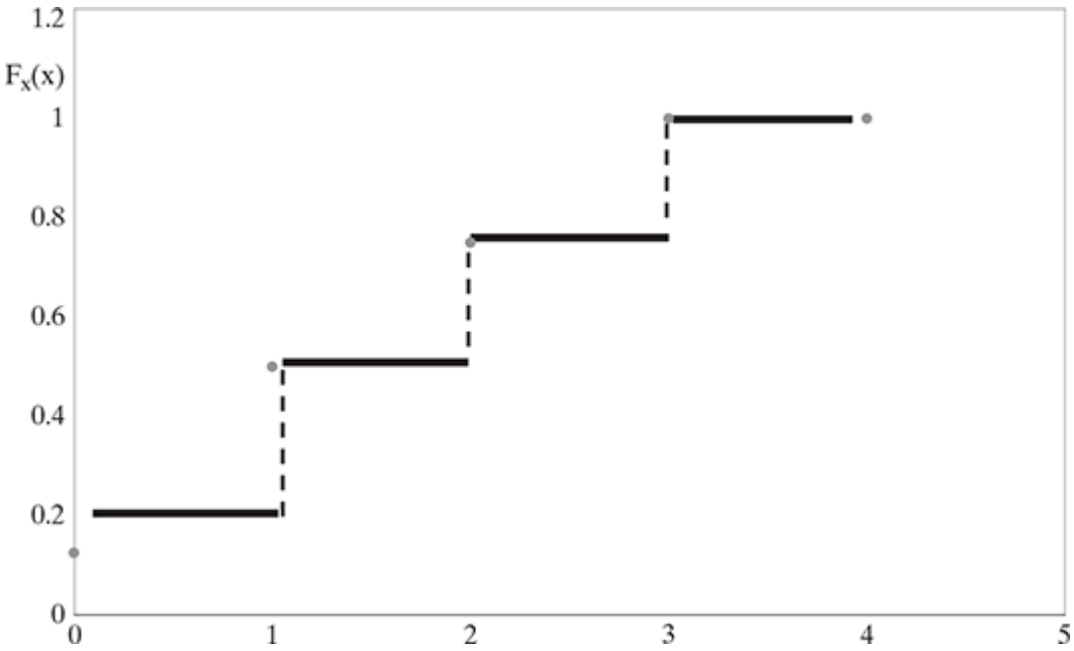


**FIG. 5.2** A discrete random variable X with cdf $F_x(x) = p(X \leq x)$ for x = 0, 1, 2, 3, 4, and $F_x(x)$ has jumps at x = 0, 1, 2, 3, 4

This satisfies the properties $0 \leq p(x) \leq 1$ and $\Sigma_{x \in X}\, p(x) = 1$.

The cumulative distribution function (cdf) $F_x(x)$ of a discrete random variable $X$ can be denoted as

$$F_x(x) = P(X \leq x) = \Sigma_{x \in X}\, p(x) \qquad (5.13)$$

Refer to Figure 5.3, the pmf is defined on a finite **state space** $X = \{1, 2, 3, 4\}$ that shows a uniform distribution with $p(x) = \frac{1}{2}$. This distribution means that $X$ is always equal to the value $\frac{1}{2}$, in other words, it is a constant.

**Example**. Suppose $X$ is a discrete random variable with $R_{X\in}\{0, 1, 2, ...\}$. we can prove .

From the equation above,

$$P(X > 0) = P_X(1) + P_X(2) + P_X(3) + P_X(4) + ...,$$
$$P(X > 1) = P_X(2) + P_X(3) + P_X(4) + ...,$$
$$P(X > 2) = P_X(3) + P_X(4) + P_X(5) + ....$$

Thus,

**FIG. 5.3** A uniform distribution within the set {1, 2, 3, 4} where $p(x = k) = \frac{1}{2}$

**DISCUSSION POINTS**

Discuss few examples of discrete random variables in practical life. How do you think the behaviour is different from the continuous random variables?

### 5.4.2 Continuous random variables

We have discussed the probabilities of uncertain discrete quantities. But most of the real-life events are continuous in nature. For example, if we have to measure the actual time taken to finish an activity then there can be an infinite number of possible ways to complete the activity and thus the measurement is continuous and not discrete as it is not similar to the discrete event of rolling a dice or flipping a coin. Here, the function $F_x(x)$

describing the event is continuous and also has a derivative that

exists and is piecewise continuous. The **probability density function (pdf)** of the continuous random variable x is defined as

We will now see how to extend probability to reason about uncertain continuous quantities. The probability that $x$ lies in any interval $a \leq x \leq b$ can be computed as follows.

We can define the events $A = (x \leq a)$, $B = (x \leq b)$, and $W = (a < x \leq b)$. We have that $B = A \cup W$, and since $A$ and $W$ are mutually exclusive, according to the sum rules:

and hence,

Define the function $F(q) = p(X \leq q)$. The **cumulative distribution function (cdf)** of random variable $X$ can be obtained by

which is a monotonically increasing function. Using this notation, we have

using the pdf of $x$, we can compute the probability of the continuous variable being in a finite interval as follows:

As the size of the interval gets smaller, it is possible to write

#### 5.4.2.1 Mean and variance

The mean in statistical terms represents the weighted average (often called as an expected value) of the all the possible values of random variable $X$ and each value is weighted by its probability. It is denoted by $\mu_x$ or $E(X)$ and defined as

mean is one of the important parameters for a probability distribution.

Variance of a random variable $X$ measures the spread or dispersion of $X$. If $E(X)$ is the mean of the random variable $X$, then the variance is given by

**Points to Ponder:**

The difference between Probability Density Function (pdf) and the Probability Mass Function (pmf) is that the latter is associated with the continuous random variable and the former is associated with the discrete random variable. While pmf represents the probability that a discrete random variable is exactly equal to some value, the pdf does not represent a probability by itself. The integration of pdf over a continuous interval yields the probability.

## 5.5 SOME COMMON DISCRETE DISTRIBUTIONS

In this section, we will discuss some commonly used parametric distributions defined on discrete state spaces, both finite and countably infinite.

### 5.5.1 Bernoulli distributions

When we have a situation where the outcome of a trial is withered 'success' or 'failure', then the behaviour of the random variable $X$ can be represented by Bernoulli distribution. Such trials or experiments are termed as Bernoulli trials.

So we can derive that, a random variable $X$ is called Bernoulli random variable with parameter p when its pmf takes the form of

$$Px\ (k) = P(X = k) = p^{k}(1\ 9\ p\ )^{1\ -k}$$

Where $0 \leq p \leq 1$. So, using the cdf $F_x(x)$ of Bernoulli random variable is expressed as

The mean and variance of Bernoulli random variable $X$ are

here, the probability of success is p and probability of failure is $1 - p$. This is obviously just a special case of a Binomial distribution with $n = 1$ as we will discuss below.

### 5.5.2 Binomial distribution

If n independent Bernoulli trials are performed and X represents the number of success in those n trials, then X is called a binomial random variable. That's the reason a Bernoulli random variable is a special case of binomial random variable with parameters $(1, p)$.

The pmf of $X$ with parameters $(n, p)$ is given by

Where, $0 \leq p \leq 1$ and

is also called the binomial coefficient which is the number of ways to choose $k$ items from $n$.

For example, if we toss a coin $n$ times. Let $X \in \{0, ..., n\}$ be the number of heads. If the probability of heads is $p$, then we say $X$ has a **binomial** distribution, written as $X \sim \text{Bin}(n, p)$.

The corresponding cdf of $x$ is

This distribution has the following mean and variance:

Figure 5.4 shows the binomial distribution of $n = 6$ and $p = 0.6$

**FIG. 5.4** A binomial distribution of n = 6 and p = 0.6

### 5.5.3 The multinomial and multinoulli distributions

The binomial distribution can be used to model the outcomes of coin tosses, or for experiments where the outcome can be either success or failure. But to model the outcomes of tossing a $K$-sided die, or for experiments where the outcome can be multiple, we can use the **multinomial** distribution. This is defined as: let $\mathbf{x} = (x_1, \ldots, x_K)$ be a random vector,

where $x_j$ is the number of times side $j$ of the die occurs. Then $\mathbf{x}$ has the following pmf:

where, $x_1 + x_2 + \ldots . x_K = n$ and

the multinomial coefficient is defined as

and the summation is over the set of all non-negative integers $x_1, x_2, \ldots . x_k$ whose sum is $n$.

Now consider a special case of $n = 1$, which is like rolling a $K$-sided dice once, so $\mathbf{x}$ will be a vector of 0s and 1s (a bit vector), in which only one bit can be turned on. This means, if the dice shows up face $k$, then the $k$'th bit will be on. We can consider $x$ as being a scalar categorical random variable with $K$ states or values, and $\mathbf{x}$ is its **dummy encoding** with $\mathbf{x} = [\Pi(x = 1), \ldots, \Pi(x = K)]$.

For example, if $K = 3$, this states that 1, 2, and 3 can be encoded as (1, 0, 0), (0, 1, 0), and (0, 0, 1). This is also called a **one-hot encoding**, as we interpret that only one of the $K$ 'wires' is 'hot' or on. This very common special case is known as a **categorical** or **discrete** distribution and because of the analogy with the Binomial/ Bernoulli distinction, Gustavo Lacerda suggested that this is called the **multinoulli distribution.**

### 5.5.4 Poisson distribution

Poisson random variable has a wide range of application as it may be used as an approximation for binomial with parameter (n,p) when n is large and p is small and thus np is of moderate size. An example application area is, if a fax machine has a faulty transmission line, then the prob-

ability of receiving an erroneous digit within a certain page transmitted can be calculated using a Poisson random variable.

So, a random variable $X$ is called a Poisson random variable with parameter $\lambda$ (>0) when the pmf looks like

the cdf is given as:

Figure 5.5 shows the Poisson distribution with $\lambda = 2$

**FIG. 5.5** A Poisson distribution with $\lambda = 2$

## 5.6 SOME COMMON CONTINUOUS DISTRIBUTIONS

In this section we present some commonly used univariate (one-dimensional) continuous probability distributions.

### 5.6.1 Uniform distribution

The pdf of a uniform random variable is given by:

And the cdf of $X$ is:

See Figure 5.6 that represents a uniform distribution with a = 3, b = 8 with increment 0.5 and repetition 20.

**FIG. 5.6** A uniform distribution with a = 3, b = 8 with increment 0.5 and repetition 20

**Example.** If X is a continuous random variable with X~Uniform(a,b). Find E(X). We know as per equation 5.33

This is also corroborated by the fact that because X is uniformly distributed over the interval $[a, b]$ we can expect the mean to be at the middle point, $E(X) =$

The mean and variance of a uniform random variable X are

This is a useful distribution when we don't have any prior knowledge of the actual pdf and all continuous values in the same range seems to be equally likely.

### 5.6.2 Gaussian (normal) distribution

The most widely used distribution in statistics and machine learning is the Gaussian or normal distribution. Its pdf is given by

and the corresponding cdf looks like

For easier reference a function $\Phi(z)$ is defined as

Which will help us evaluate the value of $F_x(x)$ which can be written as

It can be derived that

Figure 5.7 shows the normal distribution graph with mean = 4, s.d. = 12.

**FIG. 5.7** A normal distribution graph

For a normal random variable, mean and variance are

the notation $N(\mu; \sigma^2)$ is used to denote that $p(X = x) = N(x \mid \mu; \sigma^2)$

A standard normal random variable is defined as the one whose mean is 0 and variance is 1 which means $Z = N(0;1)$. The plot for this is sometimes called the bell curve, (see Fig 5.8)

The Gaussian or Normal distribution is the most widely used distribution in the study of random phenomena in nature statistics due to few reasons:

1. It has two parameters that are easy to interpret, and which capture some of the most basic properties of a distribution, namely its mean

and variance.

2. The central limit theorem ( Section 5.8) provides the result that sums of independent random variables have an approximately Gaussian distribution, which makes it a good choice for modelling residual errors or ‘noise’.
3. The Gaussian distribution makes the least number of assumptions, subject to the constraint of having a specified mean and variance and thus is a good default choice in many cases.
4. Its simple mathematical form is easy to implement, but often highly effective.

**FIG. 5.8** A standard normal distribution graph

### 5.6.3 The laplace distribution

Another distribution with heavy tails is the **Laplace distribution,** which is also known as the **double-sided exponential** distribution. This has the

following pdf:

Here $\mu$ is a location parameter and $b$ >0 is a scale parameter.

Figure 5.9 represents the distribution of Laplace

**FIG. 5.9** The distribution graph of Laplace

This puts more probability density at 0 than the Gaussian. This property is a useful way to encourage sparsity in a model.

## 5.7 MULTIPLE RANDOM VARIABLES

Till now we were dealing with one random variable in a sample space, but in most of the practical purposes there are two or more random variables on the same sample space. We will discuss their associated distribution and interdependence.

### 5.7.1 Bivariate random variables

Let us consider two random variables $X$ and $Y$ in the sample space of $S$ of a random experiment. Then the pair $(X, Y)$ is called s bivariate random variable or two-dimensional random vector where each of $X$ and $Y$ are associated with a real number for every element of $S$. The range space of bivariate random variable $(X, Y)$ is denoted by $R_{xy}$ and $(X, Y)$ can be considered as a function that to each point $\zeta$ in S assigns a point $(x, y)$ in the plane.

$(X, Y)$ is called a discrete bivariate random variable if the random variables $X$ and $Y$ both by themselves are discrete. Similarly, $(X, Y)$ is called a continuous bivariate random variable if the random variables $X$ and $Y$ both are continuous and is called a mixed bivariate random variable if one of $X$ and $Y$ is discrete and the other is continuous.

### 5.7.2 Joint distribution functions

The joint cumulative distribution function (or joint cdf) of X and Y is defined as:

For the event $(X \leq x, Y \leq y)$, we can define

A = $\{\zeta \in S; X(\zeta) \leq x\}$ and $B = \{\zeta \in S; Y(\zeta) \leq y\}$

And $P(A) = F_X(x)$ and $P(B) = F_Y(y)$

Then, $F_{XY}(x, y) = P(A \cap B)$.

For certain values of $x$ and $y$, if $A$ and $B$ are independent events of $S$, then

Few important properties of joint cdf of two random variables which are similar to that of the cdf of single random variable are

1. $0 \leq F_{XY}(x, y) \leq 1$
   If $x_1 \leq x_2$ and $y_1 \leq y_2$, then
   $F_{XY}(x_1, y_1) \leq F_{XY}(x_2, y_1) \leq F_{XY}(x_2, y_2)$
2. $F_{XY}(x_1, y_1) \leq F_{XY}(x_1, y_2) \leq F_{XY}(x_2, y_2)$
3. 
4. 
5. 
6. $P(x_1 < X \leq x_2, Y \leq y) = F_{XY}(x_2, y) - F_{XY}(x_1, y)$
   $P(X < x, y_1 < Y \leq y_2) = F_{XY}(x, y_2) - F_{XY}(x, y_1)$

### 5.7.3 Joint probability mass functions

For the discrete bivariate random variable (X, Y) if it takes the values $(x_i, y_j)$ for certain allowable integers I and j, then the joint probability mass function (joint pmf) of $(X, Y)$ is given by

Few important properties of $p_{XY}(x_i, y_j)$ are

1. $0 \leq p_{XY}(x_i, y_j) \leq 1$
2. 
3. $P[(X, Y) \in A] = \Sigma (x_i, y_j) \in R_A \Sigma\, p_{XY}(x_i, y_j)$, where the summation is done over the points $(x_i, y_j)$ in the range space $R_A$.

The joint cdf of a discrete bivariate random variable $(X, Y)$ is given by

### 5.7.4 Joint probability density functions

In case $(X, Y)$ is a continuous bivariate random variable with cdf $F_{XY}(x, y)$ and then the function,

is called the joint probability density function (joint pdf) of $(X, Y)$. Thus, integrating, we get

Few important properties of $f_{xy}(x, y)$ are

1. $f_{xy}(x, y) \geq 0$
2.
3.

### 5.7.5 Conditional distributions

While working with a discrete bivariate random variable, it is important to deduce the conditional probability function as $X$ and $Y$ are related in the finite space. Based on the joint pmf of $(X, Y)$ the conditional pmf of $Y$ when $X = x_i$ is defined as

Note few important properties of $P_{Y|X}(y_j \mid x_i)$:

1. $0 \leq P_{Y|X}(y_j \mid x_i) \leq 1$
2.

In the same way, when $(X,Y)$ is a continuous bivariate random variable and has joint pdf $f_{XY}(x, y)$, then the conditional cdf of $Y$ in case $X = x$ is defined as

Note few important properties of $f_{Y|X}(y \mid x)$:

1. $f_{Y|X}(y \mid x) \geq 0$
2.

### 5.7.6 Covariance and correlation

The **covariance** between two random variables $X$ and $Y$ measure the degree to which $X$ and $Y$ are (linearly) related, which means how X varies with Y and vice versa.

So, if the variance is the measure of how a random variable varies with itself, then the covariance is the measure of how two random variables vary with each other.

Covariance can be between 0 and infinity. Sometimes, it is more convenient to work with a normalized measure, because covariance alone may not have enough information about the relationship among the random variables. For example, let's define 3 different random variables based on flipping of a coin:

Just by looking into these random variables we can understand that they are essentially the same just a constant multiplied at their output. But the covariance of them will be very different when calculating with the equation 5.55:

$$Cov(X, Y) = 2.5, Cov(X, Z) = 25, Cov\ (Y, Z) = 250$$

To solve this problem, it is necessary to add a normalizing term that provides this intelligence:

If Cov (X,Y) = 0, then we can say X and Y are uncorrelated. Then from equation 5.55, X and Y are uncorrelated if

$$E(X, Y) = E(X)E(Y)$$

If X and Y are independent then it can be shown that they are uncorrelated, but note that the converse is not true in general.

So, we should remember

- The outcomes of a Random Variable weighted by their probability is Expectation, E(X).
- The difference between Expectation of a squared Random Variable and the Expectation of that Random Variable squared is Variance: $E(X^2) - E(X)^2$.
- Covariance, E(XY) − E(X)E(Y) is the same as Variance, but here two Random Variables are compared, rather than a single Random Variable against itself.
- Correlation, Corr (X, Y) = is the Covariance

  normalized.

Few important properties of correlation are

1. $-1 \leq \text{corr}\,[X, Y] \leq 1$. Hence, in a correlation matrix, each entry on the diagonal is 1, and the other entries are between -1 and 1.
2. corr $[X, Y] = 1$ if and only if $Y = aX + b$ for some parameters $a$ and $b$, i.e., if there is a *linear* relationship between $X$ and $Y$.
3. From property 2 it may seem that the correlation coefficient is related to the slope of the regression line, i.e., the coefficient $a$ in the expression $Y = aX + b$. However, the regression coefficient is in fact given by $a$

= cov[*X*, *Y*] /var [*X*]. A better way to interpret the correlation coefficient is as a degree of linearity.

## 5.8 CENTRAL LIMIT THEOREM

This is one of the most important theorems in probability theory. It states that if $X_1$,... $X_n$ is a sequence of independent identically distributed random variables and each having mean $\mu$ and variance $\sigma^2$ and

As $n \to \infty$ (meaning for a very large but finite set) then $Z_n$ tends to the standard normal.

Or

where, $\Phi(z)$ is the cdf of a standard normal random variable.

So, according to the central limit theorem, irrespective of the distribution of the individual $X_i$'s the distribution of the sum $S_n - X_1 + ...X_n$ is approximately normal for large n. This is the very important result as whenever we need to deal with a random variable which is the sum of a large number of random variables than using central limit theorem we can assume that this sum is normally distributed.

## 5.9 SAMPLING DISTRIBUTIONS

As we discussed earlier in this chapter, an important application of statistics in machine learning is how to draw a conclusion about a set or population based on the probability model of random samples of the set. For example, based on the malignancy sample test results of some random tumour cases we want to estimate the proportion of all tumours which are malignant and thus advise the doctors on the requirement or non-requirement of biopsy on each tumour case. As we can understand, different random samples may give different estimates, if we can get some knowledge about the variability of all possible estimates derived from the random samples, then we should be able to arrive at reasonable conclusions. Some of the terminologies used in this section are defined below:

*Population* is a finite set of objects being investigated.

*Random sample* refers to a sample of objects drawn from a population in a way that every member of the population has the same chance of being chosen.

*Sampling distribution* refers to the probability distribution of a random variable defined in a space of random samples.

### 5.9.1 Sampling with replacement

While choosing the samples from the population if each object chosen is returned to the population before the next object is chosen, then it is called the sampling with replacement. In this case, repetitions are allowed. That means, if the sample size n is chosen from the population size of N, then the number of such samples is

$$N \times N \times cc. \times N = N^{\mathbf{n}}, \text{ because each object can be repeated.}$$

Also, the probability of each sample being chosen is the same and is .

For example, let's choose a random sample of 2 patients from a population of 3 patients {A, B, C} and replacement is allowed. There can be 9 such ordered pairs, like:

$$(A, A), (A, B), (A, C), (B, A), (B, B), (B, C), (C, A), (C, B), (C, C)$$

That means the number of random samples of 2 from the population of 3 is

$$N^n = 3^2 = 9$$

and each of the random sample has probability of being chosen.

### 5.9.2 Sampling without replacement

In case, we don't return the object being chosen to the population before choosing the next object, then the random sample of size n is defined as the unordered subset of n objects from the population and called sampling without replacement. The number of such samples that can be drawn from the population size of N is

In our previous example, the unordered sample of 2 that can be created from the population of 3 patients when replacement is not allowed is

(A, B), (A, C), (B, C)

Also, each of these 3 samples of size 2 has the probability of getting chosen as

### 5.9.3 Mean and variance of sample

Let us consider X as a random variable with mean $\mu$ and standard deviation $\sigma$ from a population N. A random sample of size n, drawn without replacement will generate n values $x_1$, $x_2$.....,$x_n$ for X. When samples are drawn with replacement, these values are independent of each other and can be considered as values of n independent random variables $X_1$, $X_2$,...., $X_n$, each having mean $\mu$ and variance $\sigma^2$. The sample mean is a random variable $\dot{X}$ as

As $\bar{X}$ is a random variable, it also has a mean $\mu_{\dot{x}}$ and variance $\sigma_{\dot{x}}^2$ and it is related to population parameters as:

for a large number of n, if X is approximately normally distributed, then $\bar{X}$ will also be normally distributed.

When the samples are drawn without replacement, then the sample values $x_1$, $x_2$.....,$x_n$ for random variable X are not independent. The sample mean $\bar{X}$ has the mean $\mu_{\dot{X}}$ and variance $\sigma_{\dot{X}}^2$ given by:

where, N is the size of the population and n < N. Also if X is normally distributed then $\bar{X}$ also has a normal distribution.

Now, based on the central limit theorem, if the sample size of a random variable based on a finite population is large then the sample mean is approximately normally distributed irrespective of the distribution of the population. For most of the practical applications of sampling, when the sample size is large enough (as a rule of thumb ≥ 30) the sample mean is approximately normally distributed. Also, when a random sample is drawn from a large population, it can be assumed that the values $x_1$, $x_2$.....,$x_n$ are independent. This assumption of independence is one of the key to application of probability theory in statistical inference. We use the terms like 'population is much larger than the sample size' or 'population is large compared to its sample size', etc. to denote that the population is large enough to make the samples independent of each other. In practice,

if , then independence may be assumed.

## 5.10 HYPOTHESIS TESTING

While dealing with random variables a common situation is when we have to make certain decisions or choices based on the observations or data which are random in nature. The solutions for dealing with these situations is called decision theory or hypothesis testing and it is a widely used process in real life situations. As we discussed earlier, the key component of machine learning is to use a sample- based training data which can be used to represent the larger set of actual data and it is important to estimate how confidently an outcome can be related to the behaviour of the training data so that the decisions on the actual data can be made. So, hypothesis testing is an integral part of machine learning.

In terms of statistics, a hypothesis is an assumption about the probability law of the random variables. Take, e.g. a random sample ($X_1$,..... $X_n$) of a random variable whose pdf on parameter κ is given by f(x, κ) = f($x_1$, $x_2$...., $x_n$; κ). We want to test the assumption κ = $\kappa_0$ against the assumption

$\kappa = \kappa_1$. In this case, the assumption $\kappa = \kappa_0$ is called null hypothesis and is denoted by $H_0$. Assumption $\kappa = \kappa_1$ is called alternate hypothesis and is denoted by $H_1$.

$$H_0 : \kappa = \kappa_1$$
$$H_1 : \kappa = \kappa_1$$

A simple hypothesis is the one where all the parameters are specified with an exact value, like $H_0$ or $H_1$ in this case. But if the parameters don't have an exact value, like $H_1 : \kappa \neq \kappa_1$ then $H_1$ is composite.

Concept of hypothesis testing is the decision process used for validating a hypothesis. We can interpret a decision process by dividing an observation space, say R into two regions – $R_0$ and $R_1$. If $x = (x_1, \ldots . x_n)$ are the set of observation, then if $x \in R_0$ the decision is in favor of $H_0$and if $x \in R_1$ then the decision is in favor of $H_1$. The region $R_0$is called acceptance region as the null hypothesis is accepted and $R_1$ is the rejection region. There are 4 possible decisions based on the two regions in observation space:

1. $H_0$ is true; accept $H_0$➔ this is a correct decision
2. $H_0$ is true; reject $H_0$ (which means accept $H_1$) ➔ this is an incorrect decision
3. $H_1$ is true; accept $H_1$➔ this is a correct decision
4. $H_1$ is true; reject $H_1$ (which means accept $H_0$) ➔ this is an incorrect decision

So, we can see there is the possibility of 2 correct and 2 incorrect decisions and the corresponding actions. The erroneous decisions can be termed as:

Type I error: reject $H_0$ (or accept $H_1$) when $H_0$ is true. The example of this situation is in a malignancy test of a tumour, a benign tumour is ac-

cepted as malignant tumour and corresponding treatment is started. This is also called Alpha error where good is interpreted as bad.

Type II error: reject $H_1$ (or accept $H_0$) when $H_1$ is true. The example of this situation is in a malignancy test of a tumour, a malignant tumour is accepted as a benign tumour and no treatment for malignancy is started. This is also called Beta error where bad is interpreted as good and can have a more devastating impact.

The probabilities of Type I and Type II errors are

$$P_I = P(D_1 \mid H_0) = P(x \in R_1; H_0)$$
$$P_{II} = P(D_0 \mid H_1) = P(x \in R_0; H_1)$$

where, $D_i$ (I = 0, 1) denotes the event that the decision is for accepting $H_i$. $P_I$ is also denoted by $\alpha$ and known as level of significance whereas $P_{II}$ is denoted by $\beta$ and is known as the power of the test. Here, $\alpha$ and $\beta$ are not independent of each other as they represent the probabilities of the event from same decision problem. So, normally a decrease in one type of error leads to an increase in another type of when the sample size is fixed. Though it is desirable to reduce both types of errors it is only possible by increasing the sample size. In all practical applications of hypothesis testing, each of the four possible outcomes and courses of actions are associated with relative importance or certain cost and thus the final goal can be to reduce the overall cost.

So, the probabilities of correct decisions are

$$P(D_0 \mid H_0) = P(x \in R_0; H_0)$$
$$P(D_1 \mid H_1) = P(x \in R_1; H_1)$$

## 5.11 MONTE CARLO APPROXIMATION

Though we discussed the distribution functions of the random variables, in practical situations it is difficult to compute them using the change of variables formula. Monte Carlo approximation provides a simple but powerful alternative to this. Let's first generate $S$ samples from the distribution, as $x1, \ldots, x_S$. For these samples, we can approximate the distribution of $f(X)$ by using the empirical distribution of $\{f(x_s)\}_{s=1}^{S}$.

In principle, Monte Carlo methods can be used to solve any problem which has a probabilistic interpretation. We know that by the law of large numbers, integrals described by the expected value of some random variable can be approximated by taking the empirical mean (or the sample mean) of independent samples of the random variable. A widely used sampler is Markov chain Monte Carlo (MCMC) sampler for parametrizing the probability distribution of a random variable. The main idea is to design a judicious Markov chain model with a prescribed stationary probability distribution. Monte Carlo techniques are now widely used in statistics and machine learning as well. Using the Monte Carlo technique, we can approximate the expected value of any function of a random variable by simply drawing samples from the population of the random variable, and then computing the arithmetic mean of the function applied to the samples, as follows:

This is also called Monte Carlo integration, and is easier to evaluate than the numerical integration which attempts to evaluate the function at a fixed grid of points. Monte Carlo evaluates the function only in places where there is a non-negligible probability.

For different functions $f()$, the approximation of some important quantities can be obtained:

- 
- 
- 
- median{ $x_1$, ..., $x_s$ } → median($X$)

## 5.12 SUMMARY

1. Frequentist interpretation of probability represents the long run frequencies of events whereas the Bayesian interpretation of probability tries to quantify the uncertainty of some event and thus focuses on information rather than repeated trials.
2. According to the Bayes rule, $p$ ($A$ | $B$) = p ($B$ | $A$) p($A$)/ p ($B$).
3. A discrete random variable is expressed with the probability mass function (pmf) $p(x) = p(X = x)$.
4. A continuous random variable is expressed with the probability density function (pdf) $p(x) = p(X = x)$.
5. The cumulative distribution function (cdf) of random variable $X$ can

   be obtained by

6. The equation to find out the mean for random variable $X$ is:

   when $X$ is discrete

   when $X$ is continuous.

7. The equation to find out the variance for random variable $X$ is

8. When we have a situation where the outcome of a trial is withered 'success' or 'failure', then the behaviour of the random variable $X$ can be represented by Bernoulli distribution.
9. If n independent Bernoulli trials are performed and $X$ represents the number of success in those n trials, then $X$ is called a binomial random variable.
10. To model the outcomes of tossing a $K$-sided die, or for experiments where the outcome can be multiple, we can use the multinomial distribution.
11. The Poisson random variable has a wide range of application as it may be used as an approximation for binomial with parameter ($n$, $p$) when $n$ is large and $p$ is small and thus $np$ is of moderate size.
12. Uniform distribution, Gaussian (normal) distribution, Laplace distribution are examples for few important continuous random distributions.
13. If there are two random variables $X$ and $Y$ in the sample space of $S$ of a random experiment, then the pair ($X$, $Y$) is called as bivariate random variable.
14. The covariance between two random variables $X$ and $Y$ measures the degree to which $X$ and $Y$ are (linearly) related.
15. According to the central limit theorem, irrespective of the distribution of the individual $X_i$'s the distribution of the sum $S_n - X_1 + \ldots X_n$ is approximately normal for large n.
16. In hypothesis testing:

    Type I error: reject $H_0$ (or accept $H_1$) when $H_0$ is true.
    Type II error: reject $H_1$ (or accept $H_0$) when $H_1$ is true.

17. Using the Monte Carlo technique, we can approximate the expected value of any function of a random variable by simply drawing samples from the population of the random variable, and then computing the arithmetic mean of the function applied to the samples.

## SAMPLE QUESTIONS

### MULTIPLE - CHOICE QUESTIONS (1 MARK EACH)

1. The probabilistic approach used in machine learning is closely related to:
    1. Statistics
    2. Physics
    3. Mathematics
    4. Psychology
2. This type of interpretation of probability tries to quantify the uncertainty of some event and thus focuses on information rather than repeated trials.
    1. Frequency interpretation of probability
    2. Gaussian interpretation of probability
    3. Machine learning interpretation of probability
    4. Bayesian interpretation of probability
3. The reason the Bayesian interpretation can be used to model the uncertainty of events is that it does not expect the long run frequencies of the events to happen.
    1. True
    2. False
4. $p(A,B) = p(A \cap B) = p(A \mid B)\, p(B)$ is referred as:
    1. Conditional probability
    2. Unconditional probability
    3. Bayes rule
    4. Product rule
5. Based on this **joint distribution** on two events $p(A,B)$, we can define the **this distribution** as follows:
6. $p(A) = p(A,B) = p(A \mid B = b)\, p(B = b)$

1. Conditional distribution
2. Marginal distribution
3. Bayes distribution
4. Normal distribution

7. We can define this probability as $p(A \mid B) = p(A,B)/p(B)$ if $p(B) > 0$
   1. Conditional probability
   2. Marginal probability
   3. Bayes probability
   4. Normal probability
8. In statistical terms, this represents the weighted average score.
   1. Variance
   2. Mean
   3. Median
   4. More
9. The **covariance** between two random variables $X$ and $Y$ measures the degree to which $X$ and $Y$ are (linearly) related, which means how $X$ varies with $Y$ and vice versa. What is the formula for Cov $(X,Y)$?
   1. $\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$
   2. $\text{Cov}(X,Y) = E(XY) + E(X)E(Y)$
   3. $\text{Cov}(X,Y) = E(XY)/E(X)E(Y)$
   4. $\text{Cov}(X,Y) = E(X)E(Y)/E(XY)$
10. The binomial distribution can be used to model the outcomes of coin tosses.
    1. True
    2. False
11. Two events A and B are called mutually exclusive if they can happen together.
    1. True
    2. False

## SHORT-ANSWER TYPE QUESTIONS (5 MARKS EACH)

1. Define the Bayesian interpretation of probability.
2. Define probability of a union of two events with equation.
3. What is joint probability? What is its formula?

4. What is chain rule of probability?
5. What is conditional probability means? What is the formula of it?
6. What are continuous random variables?
7. What are Bernoulli distributions? What is the formula of it?
8. What is binomial distribution? What is the formula?
9. What is Poisson distribution? What is the formula?
10. Define covariance.
11. Define correlation
12. Define sampling with replacement. Give example.
13. What is sampling without replacement? Give example.
14. What is hypothesis? Give example.

**LONG-ANSWER TYPE QUESTIONS (10 MARKS QUESTIONS)**

1. Let X be a discrete random variable with the following PMF

    1. Find the range $R_X$ of the random variable $X$.
    2. Find P($X \leq 0.5$)
    3. Find P($0.25<X<0.75$)
    4. Find P($X = 0.2 \mid X<0.6$)
2. Two equal and fair dice are rolled and we observed two numbers $X$ and $Y$.
    1. Find $R_X$, $R_Y$ and the PMFs of $X$ and $Y$.
    2. Find $P(X = 2, Y = 6)$.
    3. Find $P(X>3 \mid Y = 2)$.
    4. If $Z = X + Y$. Find the range and PMF of $Z$.
    5. Find $P(X = 4 \mid Z = 8)$.
3. In an exam, there were 20 multiple-choice questions. Each question had 44 possible options. A student knew the answer to 10 questions,

but the other 10 questions were unknown to him and he chose answers randomly. If the score of the student $X$ is equal to the total number of correct answers, then find out the PMF of $X$. What is P($X$>15)?

4. The number of students arriving at a college between a time interval is a Poisson random variable. On an average 10 students arrive per hour. Let $X$ be the number of students arriving from 10 am to 11:30 am. What is P(10<$X$≤15)?
5. If we have two independent random variables $X$ and $Y$ such that $X$~Poisson(α) and Y~Poisson(β). Define a new random variable as $Z = X + Y$. Find out the PMF of $Z$.
6. There is a discrete random variable X with the pmf.

If we define a new random variable $Y = (X + 1)^2$ then

1. Find the range of $Y$.
2. Find the pmf of $Y$.

7. If X is a continuous random variable with PDF

1.

2. Find EX and Var(X).
3. Find $P(X \geq \frac{1}{2})$.

8. If $X$ is a continuous random variable with pdf

9. If $X$~Uniform and $Y = \sin(X)$, then find $f_Y(y)$.

10. If X is a random variable with CDF

    1. What kind of random variable is $X$: discrete, continuous, or mixed?
    2. Find the PDF of $X$, $f_X(x)$.
    3. Find $E(e^X)$.
    4. Find $P(X = 0 \mid X \leq 0.5)$.

11. There are two random variables $X$ and $Y$ with joint PMF given in Table below

    1. Find $P(X \leq 2, Y \leq 4)$.
    2. Find the marginal PMFs of $X$ and $Y$.
    3. Find $P(Y = 2 \mid X = 1)$.
    4. Are $X$ and $Y$ independent?

12. There is a box containing 40 white shirts and 60 black shirts. If we choose 10 shirts (without replacement) at random, then find the joint PMF of $X$ and $Y$ where $X$ is the number of white shirts and $Y$ is the number of black shirts.

13. If $X$ and $Y$ are two jointly continuous random variables with joint PDF

    1. Find $f_X(x)$ and $f_Y(y)$.
    2. Are $X$ and $Y$ independent to each other?
    3. Find the conditional PDF of $X$ given $Y = y$, $f_{X \mid Y}(x \mid y)$.
    4. Find $E[X \mid Y = y]$, for $0 \leq y \leq 1$.

5. Find Var($X \mid Y = y$), for $0 \leq y \leq 1$.

14. There are 100 men on a ship. If $X_i$ is the weight of the ith man on the ship and $X$i's are independent and identically distributed and also $EX_i = \mu = 170$ and $\sigma_{Xi} = \sigma = 30$. Find the probability that the total weight of the men on the ship exceeds 18,000.
15. Let $X_1, X_2, \ldots\ldots, X_{25}$ are the independent and identically distributed.

    And have the following PMF

    If $Y = X_1 + X_2 + \ldots + X_n$, estimate $P(4 \leq Y \leq 6)$ using central limit theorem.