# UNIT 1 PROBABILITY DISTRIBUTIONS

**Structure** **Page Nos.**



## 1.0 INTRODUCTION

The discipline of statistics deals with the collection, analysis and interpretation of data. Outcomes may vary even when measurements are taken under conditions that appear to be the same. Variation is a fact of life. Proper statistical methods can help us understand the inherent variability in the collected data, and facilitate the appropriate analysis of the same.

Because of this variability, uncertainty is a part of many of our decisions. In medical research, for example, interest may center on the effectiveness of a new vaccine for AIDS; an agronomist may want to decide if an increase in yield can be attributed to a new strain of wheat; a meteorologist may be interested in predicting whether it is going to rain on a particular day; an environmentalist may be interested in testing whether new controls can lead to a decrease in the pollution level; an economist's interest may lie in estimating the unemployment rate, etc. Statistics, and probabilistic foundations on which statistical methods are based, can provide the models that may be used to make decisions in these and many other situations involving uncertainties.

Any realistic model of a real world phenomenon must take into account the possibilities of randomness. That is, more often that not, the quantities we are interested in will not be predicted in advance, but rather will exhibit as inherent variation that should be taken into account by the model. Such a model is, naturally enough, referred to as a probability model.

In this unit we shall see what is a random variable and how it is defined for a particular random experiment. We shall see that there are two major types of probability distribution. We shall investigate their properties and study the different applications.

## 1.1 OBJECTIVES

After reading this unit, you should be able to

- describe events and sample spaces associated with an experiment;
- define a random variable associated with an experiment;
- decide whether a random variable is discrete or continuous;
- describe the following distributions
  a) Binomial distribution
  b) Poisson distribution
  c) Uniform distribution
  d) Exponential distribution
  e) Normal distribution
  f) Chi-square distribution.

## 1.2 RANDOM VARIABLES

Definition A ”Random experiment” or a “Statistical experiment” is any act whose outcome can not be predicted in advance. Any outcome of a random experiment is known as “event”

We will start with the following illustrations :

1) The number of telephone calls received by Monica, a telephone operator in a call center in Delhi, between say 1:00 am and 3:00 am in the early morning.
2) The amount of rainfall in Mumbai on August $1^{st}$.
3) The number of misprints on a randomly chosen page of a particular book.
4) The final results of the 5 one-day matches between India-Pakistan.
5) The outcome of rolling die.
6) The volume of sales of a certain item in a given year.
7) Time to failure of a machine.

In all the above cases there is one common feature. These experiments describe the process of associating a number to an outcome of the experiment (i.e. to an event). A function which associates a number to each possible outcome of the experiment is called a “random variable”. It is often the case that our primary interest is in the numerical value of the random variable rather than the outcome itself. The following examples will help to make this idea clear.

EXAMPLE 1: Suppose we are interested in the number of heads, say X, obtained in three tosses of a coin.
SOLUTION
If we toss a coin three times, then the experiment has a total of eight possible outcomes, and they are as follows;

$a_1=\{HHH\}$, $a_2=\{HHT\}$, $a_3=\{HTH\}$, $a_4=\{HTT\}$
$a_5=\{THH\}$, $a_6=\{THT\}$, $a_7=\{TTH\}$, $a_8=\{TTT\}$

Denoting the event corresponding to getting k heads, k=0,1,2,3,.. as$\{X$=k$\}$, observe that

$\{X=0\} => \{a_8\}$ ; $\{X=1\} => \{a_4, a_6, a_7\}$ ; $\{X=2\} =>\{a_2, a_3, a_5\}$ ; $\{X=1\} =>\{a_1\}$

Note that each value in the support of X corresponds to some element (or set of elements) in the sample space **S.** For example the , the value 0 corresponds to the element $\{a_8\}$, while the value 1 corresponds to the set of elements $\{a_4, a_6, a_7\}$

Therefore the sample space **S,** the set of all possible outcomes of this experiment can be expressed as

$$\mathbf{S}=\{a_1, a_2, \ldots, a_8\}$$

Since X is the characteristic, which denotes the number of heads out of the three tosses, it is associated with each outcome of this experiment. Therefore, X is a function defined on the elements of **S** and the possible values of *X* are $\{0,1,2,3\}$ . The set of possible values that X can take is called the support of *X* which may be denoted as $\chi$ . Observe, that *X* can be explicitly expressed as follows;

$X(a_1)=3,\ X(a_2)=X(a_3)=X(a_5)=2,\ X(a_4)=X(a_6)=X(a_7)=1,\ X(a_8)=0$

It is interesting to observe that to each value there is always some element in sample space or a set of element in sample spaces. For, example, the set of element in sample spaces corresponding to the value '0' is the point $\{a_8\}$; for 1, the set is $\{a_4, a_6, a_7\}$, for 2, the set is $\{a_2, a_3, a_5\}$ and for 3 the point is $\{a_1\}$.

Therefore, we can easily make the following identification of events corresponding to the values associated by X. Denoting the event corresponding to '0', as {X=0}, similarly for other values, observe that

$\{X=0\}=\{a_8\};\ \{X=1\}=\{a_4, a_6, a_7\};\ \{X=2\}=\{a_2, a_3, a_5\}\ \{X=3\}=\{a_1\}$

If we assume that the coin is unbiased and the tosses have been performed independently, the probabilities of all the outcomes of the sample space are equal, that is $P(a_1)=P(a_2)=\ldots=P(a_8)=\frac{1}{8}$ . Therefore, using the probability law of disjoint events we can easily obtain

$P(\{X=0\}) = P(\{a_8\}) = \frac{1}{8}$

$P(\{X=1\}) = P(\{a_4, a_6, a_7\}) = P(\{a_4\})+P(\{a_6\})+P(\{a_7\}) = \frac{3}{8}$

$P(\{X=2\}) = P(\{a_2, a3, a5\}) = P(\{a_2\})+P(\{a_3\})+P(\{a_5\}) = \frac{3}{8}$

$P(\{X=3\}) = P(\{a_1\}) = \frac{1}{8}$

Therefore, in this case the random variable X takes four values 0,1, 2,3 with the probabilities 1/8,3/8,3/8,1/8 respectively. It is also important to observe that

$P(\{X=0\})+P(\{X=1\})+P(\{X=2\})+P(\{X=3\})=1$

It is not a coincidence, it is always true. If we add the probabilities of all possible values of a random variable it is always one.

To sum up, we say that the random variable *X*, is a real valued function defined on all the elements of a sample space of a random experiment. The random variable takes different values, and for each value there is a probability associated with it. The sum of all the probabilities of all the points in the support $\chi$ of the random variable *X* adds up to one. The following figure will demonstrate the random variable *X*.

HHH HHT HTH HTT
THH THT TTH TTT $\xrightarrow{X}$ $\{0, 1, 2, 3\}$

**Figure 1: Schematic representation of a random variable**

Because of this representation , one can define probabilities for the set of numbers(depending on the random variable) rather than working with arbitrary space and this simplifies the problem considerably. Now we consider the following example.

EXAMPLE 2: You have purchased a new battery operated wristwatch and you have inserted one new battery into it. Suppose you are interested in the following:

a) How long will it be before the first battery needs replacement ?
b) How many battery will have to be replaced during a one year period?

SOLUTION
Note that both (a) and (b) are random variables and let us discuss them one by one. In case (a), we want to find the duration of time before the battery needs to be replaced. Note that the variable takes values continuously along a line say from the time duration A to time duration B. No values in between A and B are left out. In other words there is no break in the values assumed by this random variable.

In case (b), the random variable is the number of batteries. This variable can take values 0 or 1 or 2 etc. There is no continuity, since only non-negative integer values can be assumed. So, the range of this variable is a discrete set of points. From this discussion it is clear that the random variable $X$ defined in Example 1 is also a discrete random variable. The above examples shows that the random variables can be of two types. We will distinguish between the following two types of random variables;

1. Discrete Random Variable , and
2. Continuous Random Variable.

**Check Your Progress 1**

E 1: Suppose you take a 50-question multiple-choice examination, guessing your answer, and are interested in the number of correct answers obtained. Then

(a) What is the random variable $X$ that you will consider for this situation?
(b) What is the set of possible values of $X$ in this example?
(c) What does P($X$=10) mean in this context?

Now in the next two sections we will describe the discrete and continuous random variables in detail.

# 1.3 DISCRETE RANDOM VARIABLE

In this section, we define properly a discrete random variable and mention some of its basic properties.

DEFINITION : A random variable $X$ is said to be discrete, if the total number of values X can take is finite or countably infinite(i.e the support of $X$ is either finite or countable).

The support $\chi$ of $X$ may be listed as $\{a_0, a_1, a_2, \ldots\}$. Moreover, for each value of $a_i$, there is a probability associated with it. Suppose we denote them as $\{p_0, p_1, p_2, \ldots\}$, therefore, we have $P(X= a_i)=p_i$ for I=0,1,…. From the properties of a random variable and from the probability law, we have

(a) $p_i \geq 0$ for all $i \geq 0$

(b) $\sum_{i=0}^{\infty} p_i = p_0 + p_1 + p_2 + .... = 1$

From the above discussions, it follows that there exists a function p: $\chi \rightarrow \mathbf{R}$ as follows;

$$p(a) = \begin{cases} p_i & if \quad a = a_i; \quad i = 0,1,2..... \\ 0 & \text{otherwise} \end{cases}$$

This function p is called the probability mass function (p.m.f.) of the discrete random variable X.
The collection of the pairs $\{(a_i, p_i; I=0,1,\ldots\}$ is called the probability distribution of X.

Another function which plays a very important role for any random variable is known as the cumulative distribution function (c.d.f.) or simply the distribution function of the random variable. The c.d.f. $F : \mathbf{R} \rightarrow [0,1]$ of the random variable X is defined as

$$F(b) = P(X \leq b), \text{ for } -\infty < b < \infty.$$

In other words, F(b) denotes the probability that the random variable X takes on a value which will be less than or equal to b. Some important properties of the c.d.f. F (.) are

(a) F (b) is a non-decreasing function of b.

(b) $\lim_{b \to \infty} F(b) = 1$

(c) $\lim_{b \to -\infty} F(b) = 0$

Now we clarify these concepts with the same example discussed in the previous section. Suppose X is the random variable denoting the number of heads obtained in three independent tosses of a fair coin, then the probability mass function (p.m.f ) p is the function, p: $\chi \rightarrow \mathbf{R}$, such that

$$p(0) = \frac{1}{8}, \ p(1) = p(2) = \frac{3}{8}, \ p(3) = \frac{1}{8}$$

Therefore, $p(a_i) = p_i \geq 0$, for all $a_i$ and

$$\sum_{i=0}^{3} p_i = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$$

In this case the p.m.f of the random variable by the function p and the corresponding probability distribution is the set $\left\{\left(0,\frac{1}{8}\right),\left(1,\frac{3}{8}\right),\left(2,\frac{3}{8}\right),\left(3,\frac{1}{8}\right)\right\}$.This can also be expressed in a tabular form as follows

**TABLE 1**
**PROBABILITY DISTRIBUTION OF THE NUMBER OF HEADS IN THREE INDEPENDENT TOSSES OF A FAIR COIN**

| The number of heads (X value) | Probability |
|---|---|
| 0 | $\frac{1}{8}$ |
| 1 | $\frac{3}{8}$ |
| 2 | $\frac{3}{8}$ |

| | |
|---|---|
| 3 | $\frac{1}{8}$ |

Now let us see the graphical representation of the distribution.

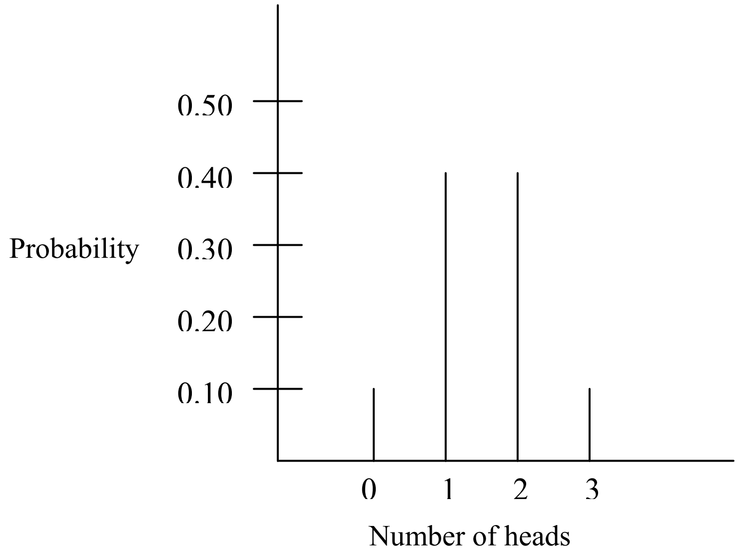



**Figure 2: Graphical representation of the distribution of $X$**

Graphically along the horizontal axis, plot the various possible values $a_i$ of a random variable and on each value erect a vertical line with height proportional to the corresponding probability $p_i$.

Now let us consider the c.d.f of the random variable X. Note that if b<0, clearly F(b)=P($X \le$ b)=0, because $X$ takes values only {0,1,2,3}. If b=0, that is F(0)=P($X \le$ 0)=P($X$=0)=1/8. If $0 \le b \le 1$, then P( $X \le$ b)=P($X$=0)+P(0<$X$<b)=1/8+0=1/8. Similarly, if b=1, F(1)= P(X ≤ 1) = P($X$=0) + P($X$=1)=1/8+3/8=4/8 and so on. Therefore, the c.d.f. F(.) has the following form;

$$F(b)=\begin{cases} 0 & if \quad b<0 \\ \frac{1}{8} & if \quad 0 \le b < 1 \\ \frac{4}{8} & if \quad 1 \le b < 2 \\ \frac{7}{8} & if \quad 2 \le b < 3 \\ 1 & if \quad b \le 3 \end{cases}$$

**Note :Mathematical expectation or Expected values or Expectations_** forms the fundamental idea in the study of probability distribution of any discrete random variable $X$ , the expected value (or mean) , denoted as E($X$) is defined as

$$E(X) = x_0p_0 + x_1p_1 + x_2p_2 + \ldots\ldots\ldots\ldots = \Sigma x_ip_i$$

Where $x_0$, $x_1$, $x_2$ etc are the values assumed by X and $p_0$, $p_1$, $p_2$ etc are probabilities of these values. Under special conditions (like all probabilities are equal )then

$$E(X) = \text{mean of } x_0, x_1, x_2 \ldots\ldots$$

Similarly for continuous variables X having density function p(x) where P[$X$=x] = p(x) , the Expectation E($X$) will be given by integral of $x_ip(x_i)$ w.r.t x
This concept of Expectation also contributes to the definition of **Moment Generating Function of *X* i.e $M_x(t)= E(e^{tx})$**

Example 3 A box contains twice as many red marbles as green marbles. One marble is drawn at random from the box and is replaced ; then a second marble is drawn at random from the box.If both marbles are green you win Rs. 50 ; if both marbles are red you loose Rs. 10 ; and if the y are of different colour then neither you loose nor you win. Determine the probability distribution for the amount you win or loose?

Solution Say X denote the amount you win (+) or loose (-) ; i.e X= +50 or -10

The probability that both marbles are green is 1/9 i.e. P[X=+50] =1/9
The probability that both marbles are red is 4/9 i.e. P[X=-10] =4/9
The probability that marbles are of different colours is 4/9 i.e. P[X=0] =4/9
Thus the probability distribution is given by following table

| **Amount(in Rs won(+) or lost(-))** | **Probability** |
|---|---|
| +50 | 1/9 |
| 0 | 4/9 |
| -10 | 4/9 |

**Check Your Progress 2**

**E1**: Which of the variables given below are discrete? Give reasons for your answer.

(a) The daily measurement of snowfall at Shimla.
(b) The number of industrial accidents in each month in West Bengal.
(c) The number of defective goods in a shipment of goods from a manufacturer.

### 1.3.1 Binomial Distribution

One very important discrete random variable (or discrete distribution) is the binomial distribution. In this subsection, we shall discuss this random variable and its probability distribution.

Quite often we have to deal with the experiments where there are only two possible outcomes. For example, when a coin is tossed either a head or a tail will comes up, a newborn is either a girl or a boy, a seed either germinates or fails to germinate. Let us consider such an experiment. For example consider the same experiment of tossing a coin independently three times. Note that the coin need not necessarily be a fair one, that is P(Head) may not be equal at P(Tail)

This particular experiment has a certain characteristic. First of all it involves repetition of three identical experiments (trials). Each trial has only two possible outcomes: a Head or a Tail. We refer to the outcome ‘Head’ as success and the outcome ‘Tail’ as failure. All trials are independent of each other. We also know that the probability of getting a ‘Head’ in a trial is p and probability of getting a ‘tail’ in a trial is 1 – p, that is

P(Head) = P(success) = p and P(Tail) = P(failure) = q = 1 - p

This shows that the probability of getting a ‘success’ or a ‘failure’ does not change from one trial to another. If X denotes the total number of ‘Heads’, obtained in three trials, then

X is a random variable, which takes values {0,1,2,3}. Then regarding the above experiment, we have observed the following;

(1) It involves a repetition of n identical trials (Here n=3).
(2) The trials are independent of each other.
(3) Each trial has two possible outcomes.
(4) The probability of success (p) and the probability of failure' (q=1-p) remain constant and do not change from trial to trial.

Now let us try to compute the probabilities P{$X$ = 0}, P{$X$ = 1}, P{$X$ = 2} and P{$X$ = 3} in this case. Note that

P($X$=0) = P(getting tails in all three trials)
= P({TTT}) = $(1\text{-}p)^3 = q^3$.

Similarly,

P($X$ = 1) = P(getting one Tail and two Heads in three trials)
= P({THH,HTH,HHT}) = P({THH}) + P({HTH}) + ({HHT})
= $(1\text{-}p)^2p+(1\text{-}p)^2p+(1\text{-}p)^2p = 3(1\text{-}p)^2p = 3q^2p$.

Similarly,

P($X$ = 2) = (getting two Tails and one Head in three trials)
= P({HTT,THT,TTH}) = P({HTT}) + P({THT}) + ({TTH})
= $(1\text{-}p)p^2 + (1\text{-}p)p^2 + (1\text{-}p)p^2 = 3(1\text{-}p)p^2=3q^2p$

Finally

P($X$=3) = P(getting Heads in three trials)
= P({HHH}) = $p^3$

Now observe that instead of n=3, in the above example we can easily compute the probability for any general n. Suppose we compute P(X = r), for 0 ≤ r ≤ n, then note that

$$P(X = r) = C(n,r)p^r(1\text{-}p)^{n\text{-}r} = C(n,r)p^rq^{n\text{-}r},$$

Where C(n,r) denotes the number of ways n places can be filled with r Heads and n-r Tails. From your school mathematics, recall that it is the number of combination of n objects taken r at a time and it can be calculated by the following formula:

$$C(n,r) = \frac{n!}{r!(n-r)!}$$

Therefore, for r = 0,1,…,n,

$$P(X = r) = \frac{n!}{r!(n-r)!}p^r q^{n-r},$$

where
n = the number of trial made
r = the number of success
p = the probability of success in a trial
q = 1-p = the probability of a failure.

Now we define the binomial distribution formally.

Let X represents the number of successes in the set of n independent identical trials. Then X is a discrete random variable taking values 0,1,…,n. The probability of the event P(X= r) is given by

$$P(X=r)=\frac{n!}{r!(n-r)!}p^r q^{n-r},\ \text{r=0,1,2,…..,n}$$

where n, r, p, q are same as defined before. Such a random variable X is called a binomial random variable and its probability distribution is called the binomial distribution. A Binomial distribution has two parameters n and p

**Check Your Progress 3**

E1 : A farmer buys a quantity of cabbage seeds from a company that claims that approximately 90% of the seeds will germinate if planted properly. If four seeds are planted, what is the probability that exactly two will germinate?

### 1.3.2 Poisson Distribution

In this subsection we will introduce another discrete distribution called 'Poisson Distribution'. First we shall describe the different situations where we can apply this Poisson Distribution.

Suppose it is the first hour at a bank in a busy Monday morning , and we are interested in the number of customers who might arrive during that hour, or during a 5-minute or a 10-minute interval in that hour. In statistical terms , we want to find the probabilities for the number of arrivals in a time interval.

To find this probability, we are making some assumptions similar to the binomial distribution.

(a) The average arrival rate at any time, remains the same over the entire first hour.
(b) The number of arrivals in a time interval does not depend on what has happened in the previous time intervals.
(c) It is extremely unlikely that more than one customer will arrive at the same time.

Under those above assumptions, we can find the required the probabilities. Suppose X is the random variable denoting the number of customers that arrive in the first hour, then

$$P(X=i)=\frac{e^{-\lambda}\lambda^i}{i!},\qquad i=0,1,2,3,...$$

Where λ (the Greek letter Lambda) denotes the average arrival rate per hour. For example, suppose we know that average number of customers that arrive in that bank during the first hour is 60 and we want to find what is the chance there will be no more than 3 customers in the first 10 minutes. Since we know that the average arrival rate per hour is 60 , if we denote λ to be the average arrival rate per 10 minutes, then $\lambda=\frac{60\times10}{60}$=10. Therefore, we can use the above formula and get

$$P(X=i)=\frac{e^{-10}10^i}{i!},\qquad i=0,1,2,3.....$$

But we want to find the probability that no more than 3 customers will be there in the first ten minutes and that is

$$P(X\le3)=P(X=0)+P(X=1)+P(X=2)+P(X=3)\qquad(1)$$

$$=e^{-10}+e^{-10}10+\frac{e^{-10}10^2}{2!}+\frac{e^{-10}10^3}{3!}\qquad(2)$$

$$\approx 0.00005 + 0.00050 + 0.00226 + 0.00757 = 0.01038 \qquad (3)$$

What does this value 0.01038 indicate? It tells us that if the arrival rates are uniform then there is only 1% chance that less than three customers will be there, or in other words, there is a 99% chance that there will be more then 3 customers in the first 10 minutes.

Similarly if we want to find out the chance that there will be no more than 3 customers in the first 5 minutes, then similarly, as above we can see that in this case $\lambda = \frac{60 \times 5}{60} = 5$.

Therefore, if Y denotes the number of customers presents in the first 5 minutes, then

$$P(Y \le 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) \qquad (4)$$

$$= e^{-5} + 5e^{-5} + \frac{e^{-5}5^2}{2!} + \frac{e^{-5}5^3}{3!} \qquad (5)$$

$$\approx 0.00674 + 0.03369 + 0.08422 + 0.14037 = 0.26502 \qquad (6)$$

From the above two examples it is clear that if we change the time unit (and hence the value of λ), the probabilities will change. The probability mass function (p.m.f ) given by

$$p(i) = P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}, \quad i = 0,1,2,3,\ldots$$

represents the Poisson probability distribution. From the series expansion of $e^{\lambda}$ , it easily follows that

$$\sum_{i=0}^{\infty} P(X = i) = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} = 1$$

as it should be.

One point that should always be kept in mind is that a random variable denoting the number of occurrences in an interval of time will follow a Poisson distribution, if the occurrences have the following characteristics:

(a) The average occurrence rate per unit time is constant.
(b) Occurrence in an interval is independent of what has happened previously.
(c) The chance that more than one occurrence will happen at the same time is negligible.

Now let us look at some situations where we can apply the Poisson distribution. Here is an example

EXAMPLE 4: Calls at a particular call center occur at an average rate of 8 calls per 10 minutes. Suppose that the operator leaves his position for a 5 minute coffee break. What is the chance that exactly one call comes in while the operator is away?

Solution: In this case the conditions (a), (b) and (c) are satisfied. Therefore if $X$ denotes the number of calls during a 5 minute interval, then $X$ is a Poisson random variable with $\lambda = \frac{8 \times 5}{10} = 4$. Therefore,

$$P(X = 1) = \frac{e^{-4}4^1}{1!} = 4e^{-4} \approx 0.073$$

That means the chance is 7.3% that the operator misses exactly one call.

**Check your progress 4**
E 1: If a bank receives on an average $\lambda = 6$ bad Cheques per day, what is the probability that it will receive 4 bad checks on any given day

# 1.4 CONTINUOUS RANDOM VARIABLE

So far we have discussed about the discrete random variables in details and we have provided two important discrete distributions namely binomial and Poisson distributions. Now in this section we will be discussing another type of random variables namely continuous random variables.

Let us look at the part (a) of Example 2. Note that we want to find the time of occurrence rather than the number of occurrences. Therefore if the random variable *X* denotes the time of occurrence of a particular event, then the random variable *X* can take any value on the positive real line or may be any value on a fixed interval say (A,B). Therefore, the random variable can take uncountably many values**.** This type of a random variable which can take uncountably many values is called a continuous random variable. For a continuous random variable *X*, the probability that *X* takes a particular value is always zero, but we can always specify the probability of *X* of any interval through a probability density function (p.d.f.). The exact details are given below.

DEFINITION: Let *X* be a continuous random variable which takes values in the interval (A,B). A real valued function $f(x)$: $\mathbf{R} \rightarrow \mathbf{R}$ is called the p.d.f of X, if
(a) $f(x) \geq 0$ and $f(x) = 0$, if $x <$ A or $x >$ B.
(b) $\int_A^B f(x)dx = 1$
(c) $P(c < X < d) = \int_c^d f(x)dx$

Now we shall see how we can use the graph of the p.d.f. of a continuous random variable to study real life problems

Example 5: Suppose the Director of a training program wants to conduct a programme to upgrade the supervisory skills of the production line supervisors. Because the programme is self-administered, supervisors require different number of hours to complete the programme. Based on a past study, it is known that the following p.d.f. shows the distribution of time spent by a candidate to complete the program. From the graph it is clear that the average time

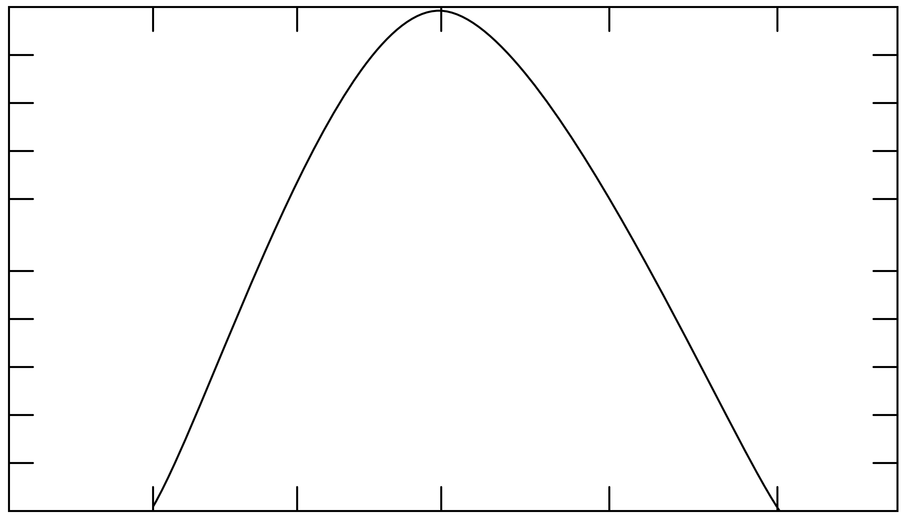



**Figure 3: The p.d.f. of the time spent by a candidate to complete the program**

Spent by the candidate is 250 and it is symmetrically distributed around 250. How can the Director use this graph to find the following. What is the chance that a participant selected at random will require

(a) more than 250 hours to complete the program
(b) less than 250 hours to complete the program

SOLUTION:Since the graph is symmetric, therefore, it is clear that area under the curve above 250 is half. Therefore, the probability that the random variable takes values higher than 250 is $\frac{1}{2}$. Similarly, the random variable takes value lower than 250 is also $\frac{1}{2}$.
Please try the following exercise now:

Now in the following subsections we will consider different continuous distributions.

### 1.4.1 The Uniform Random Variable

The uniform distribution is the simplest of a few well-known continuous distributions, which occur quite often. It can be explained intuitively very easily. Suppose X is a continuous random variable such that if we take any subinterval of the sample space, then the probability that X belongs to this subinterval is the same as the probability that X belongs to any other subintervals of the same length. The distribution corresponding to this random variable is called a uniform distribution and this random variable is called a uniform random variable.

Formally, we define the uniform random variable as follows: The random variable X is a uniform random variable between (A, B), if its p.d.f. is given by

$$f(x) = \begin{cases} \frac{1}{B-A} & for \quad A < x < B \\ 0 & \text{otherwise} \end{cases}$$

From the p.d.f. it is clear that if A < $a_1$ < $b_1$ < B, A < $a_2$ < $b_2$ <B and $b_1$-$a_1$ = $b_2$-$a_2$, then

P($a_1 < X < b_1$) = P($a_2 < X < b_2$ ). Therefore, if the length of the intervals are same then the corresponding probabilities will be also equal. Let us see some examples of such random variables:

EXAMPLE 6: A train is likely to arrive at a station at any time uniformly between 6:15 am and 6:30 am. Suppose $X$ denotes the time the train reaches, measured in minutes, after 6:00 am.

SOLUTION In this case X is a uniform random variable takes value between (15,30). Note that in this P($20 < X < 25$) is same P($18 < x < 23$) and that is equal to

$$\frac{25-30}{30-15} = \frac{23-18}{30-15} = \frac{1}{3}$$

**Check your progress 5**

E 1 An office fire drill is scheduled for a particular day, and the fire alarm is likely to ring uniformly at any time between 10:00 am to 1:00 pm.

### 1.4.2 Exponential Random Variable

In making a mathematical model for a real world phenomenon, it is always necessary to make certain simplifying assumptions so as to render the mathematical tractability. On the other hand, however, we can not make too many simplifying assumptions, for then our conclusions obtained from the mathematical model, would not be applicable to the real world problem. Thus in short, we must take enough simplifying assumptions to enable us to handle the mathematics but not so many that the mathematical model no longer resembles the real world problem. One simplifying assumption that is often made is to assume that certain random variables are exponentially distributed. The reason for this is that the exponential distribution is both easy to work and is often a good approximation to the actual distribution.

We use exponential distribution to model lifetime data that is the data, which are mainly non-negative. Although, with proper modifications, it can be used to analyze any type of data (not necessarily non-negative only). The property of the exponential distribution, which makes it easy to analyze, is that it does not deteriorate with time. By this we mean that if the lifetime of an item is exponentially distributed, then an item which has been in use for say ten hours, is as good as a new item in regards to the amount of time remaining until the item fails.

Now we define the exponential distribution formally: A continuous random variable $X$ is said to be an exponential random variable if the p.d.f of $X$ is given for some $X$>0, by

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

Here λ is known as the rate constant. It can be shown mathematically that the average value or the mean values of X is $\frac{1}{\lambda}$.Shapes of f (x) for different values of λ are provided in the figure below. From the figure,

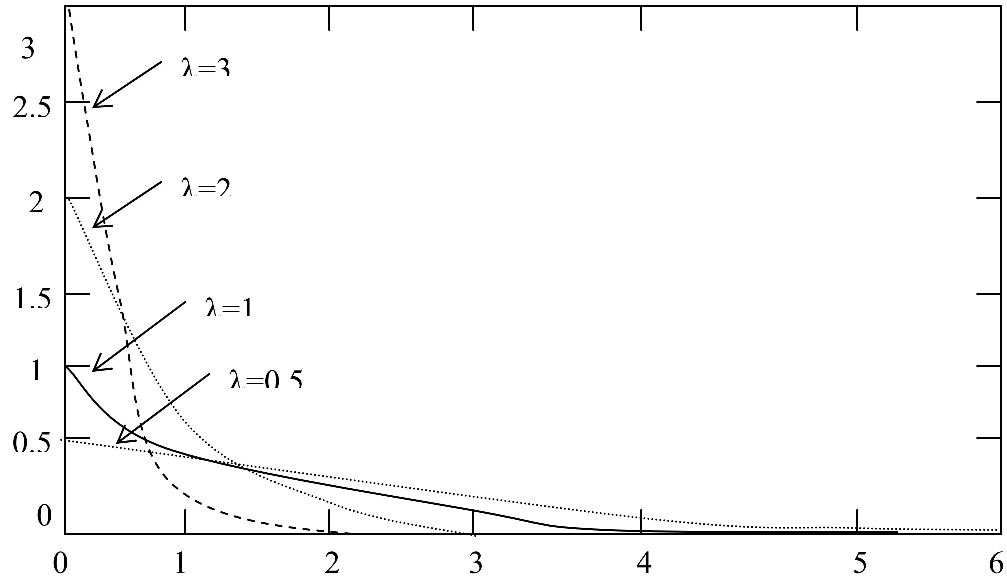



**Figure 4: The p.d.f. of the exponential distribution for different values of λ.**

It is clear that f(x) is a decreasing function for all values of λ and f(x) tends to 0 as x tends to ∞. Now consider the following example.

EXAMPLE 7: suppose that the amount of time one spends in a bank to withdraw cash from an evening counter is exponentially distributed with mean ten minutes, that is λ = 1/10. What is the probability that the customer will spend more than fifteen minutes in the counter?

SOLUTION: If X represents the amount of time that the customer spend in the counter than we need to find P(*X*>15). Therefore,

$$P\,(X > 15) = \int_{15}^{\infty} \lambda\, e^{-\lambda x} = e^{-15\lambda} = e^{-\frac{3}{2}} \approx 0.223$$

P(*X*>15)= .223 represents that there is a 22.3 % chance that the customer has to wait more than 15 minutes.

### 1.4.3 Normal Distribution

Normal distribution is undoubtedly the most used continuous distribution in different areas like astronomy, biology, psychology and of course in probability and statistics also. Because of its practical as well as theoretical importance it has received considerable attentions in different fields. The normal distribution has a unique position in probability theory, and it can be used as an approximation to other distributions. In practice 'normal theory' can frequently be applied with small risk of serious error, when substantially non-normal distributions correspond more closely to the observed value. The work of Gauss in 1809 and 1816 established techniques based on normal distribution, which became standard methods used during the nineteenth century. Because of Gauss's enormous contribution, it is popularly known as Gaussian distribution also.

We will now state the normal distribution formally: The random variable *X* is said to be normally distributed with parameters $\mu$ and $\sigma$ , if the p.d.f f(x) of *X* is given by

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

Here $\mu$ is a real number lying between $-\infty$ and $\infty$ and $\sigma$ is a real number lying between 0 and $\infty$.

The function f(x) may look a bit strange, but do not get bother. Just notice the following important things. Note that it involves two parameters $\mu$ and $\sigma$, that means corresponding to each $\mu$ and $\sigma$ we get a distribution function. More over it can be seen that for $-\infty < \mu < \infty$ and $0 < \sigma < \infty$, the function f(x) is symmetric about $\mu$ and is a 'bell shaped' one. Both $\mu$ and $\sigma$ have nice interpretation. It can be easily checked that $\mu$ is the average value or mean of the distribution and $\sigma$ provides the measure of spread. The p.d.f. of two different normal distributions are provided below.

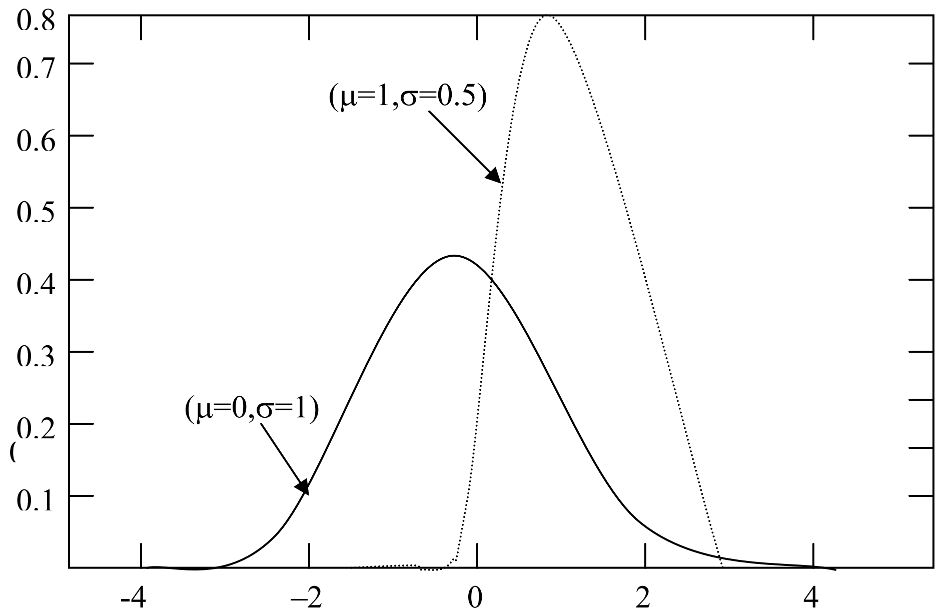



**Figure 5: The p.d.f of the normal distribution for two different values of ( μ, σ).**

It is clear from the above figure that the p.d.f. is symmetric about $\mu$ and the shape depends on $\sigma$. The spread of the distribution is more if $\sigma$ is large.

Now let us find the P(*a* < *X* < *b*) for any *a* and *b,* when *X* follows normal distribution with parameters $\mu$ and $\sigma$, note that,

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z2}{2}} dz.$$

The last equality follows by simply making the transformation $z = \frac{x-\mu}{\sigma}$. Therefore it follows

$$P(a < X < b) = P\left( \frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma} \right),$$

Where Z follows normal distribution with parameters 0 and 1. Although the probability can not be calculated in a compact form, extensive tables are available for P (Z < z) for different values of z. The table values can be used to compute P (a < X < b) for any $\mu$ and $\sigma$.

Say we denote F(a) = P[Z<= a], the probability thet the standard normal variable Z takes values less than or equal to 'a'.The values of F for different values of a are calculated and listed in table . One such table is given in the end of this unit

Note that the entries in the table are values of z for z=0.00 0.01,0.02 , ..0.09.To find the probability that a random variable having standard normal distribution will take on a value between a and b, we use the equation

$$P[a < Z < b] = F(b) – F(a)$$

And if either a or b is negative then we can make use of the identity

$$F(-z) = 1 - F(z)$$

EXAMPLE 8 Use the table to find the following probabilities

(a) P[0.87 < Z < 1.28]
(b) P[-0.34 < Z < 0.62]
(c) P[Z >= 0.85]
(d) P[Z >= -0.65]

SOLUTION

a)P[0.87 < Z < 1.28] : Find F(1.28 from the table). In the table in the row for Z=1.2 find the value under column 0.08 it will be 0.8997 . Similarly find F(0.87) =0.8078

so, P[0.87 < Z < 1.28] = 0.8997 – 0.8078 = 0.0919

b) Similarly P[-0.34 < Z < 0.62] = F(0.62) – F(0.34) = F(0.62) – [ 1- F(0.34)]
= 0.7324 – (1 - 0.6331) = 0.3655

c) Similarly calculate P[Z > 0.85] = 1- P[Z <= 0.85] = 1-F(0.85) = 0.1977

d)P[Z > -0.65] = 1 – P [ Z <= - 0.65]
= 1 – F (-0.65)
= 1- F(1-F(0.65))
= 0.7422

Next we shall see that how to use the standard normal probability table to calculate probability of any normal distribution

**Standardising**

Any normal random variable X, which has mean Φ and variance $\sigma^2$ can be standardized as follows.

Take a variable X, and

i) subtract its mean (m or Φ) and then,
ii) divide by its standard deviation(s or $\sigma$).

We will call the result, Z, so

$$Z = \frac{X - \mu}{\sigma}$$

For example, suppose, as earlier, that X is an individual's IQ score and that it has a normal distribution with mean $\Phi$ = *100* and standard deviation $\sigma = 15$. To standardize and individuals IQ score, X, we subtract $\Phi$ = 100 and divide the result by $\sigma = 15$ to give,

$$Z = \frac{X - 100}{15}$$

In this way every value of X, has a corresponding value of Z. For instance, when X = 130, $Z = \frac{130 - 100}{15} = 2$ and when X = 90, $Z = \frac{90 - 100}{15} = -0.67$.

**The distribution of standardized normal random variables**

The reason for standardizing a normal random variable in this way is that a standardized normal random variable

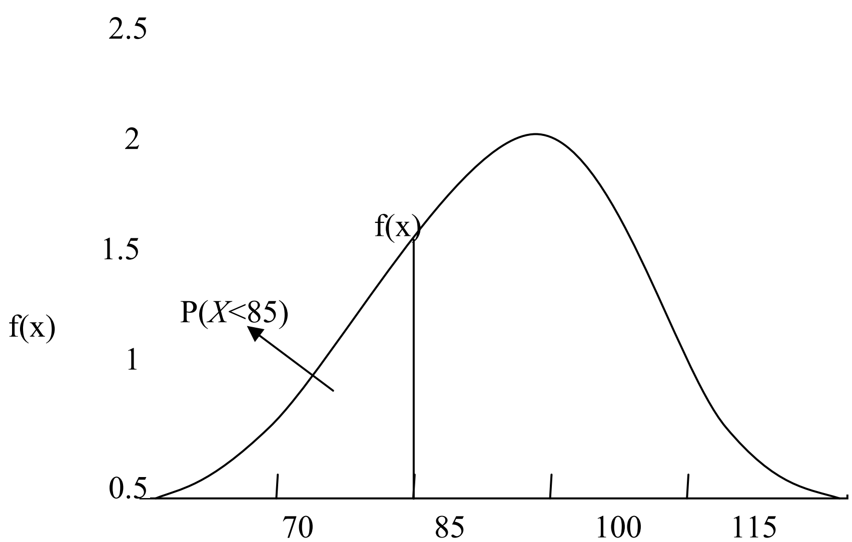
$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution.

That is, Z is N(0,1). So, if we take any normal random variable, subtract its mean and then divide by its standard deviation, the resulting random variable will have standard normal distribution. We are going to use this fact to calculate (non-standard) normal probabilities.

**Calculating probabilities**

With reference to the problem of IQ score, suppose we want tot find the probability that an individual's IQ score is less than 85, i.e. P[X<85]. The corresponding area under the pdf N(100,$15^2$) is shown in Figure 6 below.




x F
**Figure 6:** area under the pdf N(100,$15^2$)
Figure 5: Density function *f*

We cannot use normal tables directly because these give N(0,1) probabilities. Instead, we will convert the statement X<85 into an equivalent statement which involves the standardized score, $Z = \dfrac{X-100}{15}$ because we know it has a standard normal distribution.

We start with X=85. To turn X into Z we must standardize the X, but to ensure that we preserve the meaning of the statement we must treat the other side of the inequality in exactly the same way. (Otherwise we will end up calculating the probability of another statement, not X<85). 'Standardising' both sides gives,

$$\frac{X-100}{15} < \frac{85-100}{15}.$$

The left hand side is now a standard normal random variable and so we can call it Z, and we have,

$$Z < \frac{85-100}{15}$$

which is

$$Z < -1.$$

So, we have established that the statement we started with, X < 85 is equivalent to Z < – 1. This means that whenever an IQ score, X is less than 85 the corresponding standardized score, Z will be less than – 1 and so the probability we are seeking, P[X<85] is the same P[Z < – 1].

P[Z < – 1] is just a standard normal probability and so we can look it up in Table 1 in the usual way, which gives 0.1587. We get that P[X < 85] = 0.1587.

This process of rewriting a probability statement about X, in terms of Z, is not difficult if you are systematically writing down what you are doing at each stage. We would lay out the working we have just done for P[X < 85] as follows.

X has a normal distribution with mean 100 and standard deviation 15. Let us find the probability that X is less than 85.

$$P[X<85] = P\left[\frac{X-100}{15} < \frac{85-100}{15}\right]$$

$$= P[Z - 1] = 0.1587$$

Let us do some problems now.

Example 9: For each of these write down the equivalent standard normal probability.

a) The number of people who visit a historical monument in a week is normally distributed with a mean of 10,500 and a standard deviation of 600. Consider the probability that fewer than 9000 people visit in a week.
b) The number of cheques processed by a bank each day is normally distributed with a mean of 30,100 and a standard deviation of 2450. Consider the probability that the bank processes more than 32,000 cheques in a day.

Solution: Here, we want to find the standard normal probability corresponding to the probability P[X < 9000].

a) We have $P[X < 9000] = P\left[\frac{X - 10500}{600} < \frac{9000 - 10500}{600}\right] = P[Z < -2.5]$.

b) Here, we want to find the standard normal probability corresponding to the probability P[X > 32000].

$$P[X < 32000] = P\left[\frac{X - 30100}{2450} < \frac{32000 - 30100}{2450}\right] = P[Z < -0.78]$$

**Note**: Probabilities like P[a < X < b] can be calculated in the same way. The only difference is that when X is standardized, similar operations must be applied to both a and b. that is, a < X < b becomes,

$$\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}$$

which is

$$\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}$$

Example 10: An individual's IQ score has a $N(100, 15^2)$ distribution. Find the probability that an individual's IQ score is between 91 and 121.

Solution: We require P[91 < X < 121]. Standardising gives

$$P\left[\frac{91-100}{15} < \frac{X-100}{15} < \frac{121-100}{15}\right]$$

The middle term is standardized normal random variable and so we have,

$$P\left[\frac{-9}{15} < Z < \frac{21}{15}\right] = P[-0.6 < Z < 1.4] = 0.9192 - 0.2743 = 0.6449.$$

**<u>Check your progress 6</u>**

E1 If a random variable has the standard normal distribution , find the probability that it will take on a value

a) Less than 1.50
b) Less than –1.20
c) Greater than –1.75

E2 A filling machine is set to pour 952 ml of oil into bottles. The amount to fill are normally distributed with a mean of 952 ml and a standard deviation of 4 ml. Use the standard normal table to find the probability that the bottle contains oil between 952 and 956 ml ?

### 1.4.4 Chi-Square Distribution

In the last subsection we have discussed normal distribution. The chi-square distribution can be obtained from the normal distribution as follows. Suppose $Z_1,\ldots,Z_n$ are n independent identically distributed normal random variables with parameters 0 and 1, then $Z_1^2 + \ldots + Z_n^2$ is said to have chi-square distribution with n degrees of freedom. The degrees of freedom here basically indicates the number of independent components which constitute the chi-square distribution. It has received several attention because of its appearance in the constructing analysis of variable tables, contingency tables and for obtaining the critical values of different testing procedure. Now we formally provide the p.d.f of a chi-square random variable with n degrees of freedom.

If the random variable $X$ has chi-square distribution with n-degrees of freedom, then the p.d.f. of $X$ is $f(x)$

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1} & \text{if } x > 0 \\ 0 & \textit{otherwise} \end{cases}$$

here Γ (.) is a gamma function and it is defined as

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

Although, the p.d.f of chi-square random variable is not a very nice looking one, do not bother about that. Keep in mind that the shapes of density functions are always skewed. In this case also if we want to compute P(a < X < b) for any a, b and n, explicitly it is not possible to compute. Numerical integration is needed to compute this probability. Even for chi-square distribution extensive tables of P(a < X < b) are available for different values of a, b and n.

**Note :** We have a standard table corresponding to Chi- Square Distribution, many times you may need to refer the values from the table. So the same is given at the end , method of usage is similar to that discussed in Normal distribution.

EXAMPLE 9 Show that the moment generating function of a random variable $X$ which is chi-square distributed with $v$ degrees of freedom is $M(t) = (1 - 2t)^{-v/2}$.

SOLUTION $M(t) = E(e^{tx}) = \dfrac{1}{2^{v/2}\Gamma(\text{v}/2)} \displaystyle\int_0^\infty e^{tx} x^{(v-2)/2} e^{-x/2} dx$

$$= \frac{1}{2^{v/2}\Gamma(\text{v}/2)} \int_0^\infty x^{(v-2)/2} e^{-x(1-2t)x/2} dx$$

Letting $(1-2t)^{x/2} = u$ in the last integral we find

$$M(t) = \frac{1}{2^{v/2}\Gamma(\text{v}/2)} \int_0^\infty \left(\frac{2u}{1-2t}\right)^{(v-2)/2} e^{-u} \frac{2du}{1-2t}$$

$$= \frac{(1-2t)^{-v/2}}{\Gamma(v/2)} \int_0^\infty u^{(v/2)-1} e^{-u} du = (1-2t)^{-v/2}$$

**Check your progress 7**

**E1** Let $X_1$ and $X_2$ be independent random variables, which are chi-square distributed with $v_1$ and $v_2$ degrees of freedom respectively. Show that the moment generating function of $Z = X_1 + X_2$ is $(1-2t)^{-v(v_1+v_2)/2}$

**E2** Find the values of $x^2$ for which the area of the right-hand tail of the $x^2$ distribution is 0.05, if the number of degrees of freedom $v$ is equal to (a) 15, (b) 21, (c) 50.

## 1.5 SUMMARY

In this unit we have covered following points:

a) A random variable is a variable that takes different values according to the chance outcome

b) Types of random variables: Discrete and Continuous

c) Probability distribution gives the probabilities with which the random variables takes various values in their range

d) Discussed probability distributions :

   a. Binomial Distribution : The probability of an event P[$X$=r] in this distribution is given by

   $$P(X = r) = C(n,r)p^r(1-p)^{n-r} = C(n,r)p^r q^{n-r},$$

   b. Poisson Distribution : The probability of an event P[$X$=i] in this distribution is given by

   $$P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}, \qquad i = 0,1,2,3,...$$

   c. Uniform Distribution : The probability density function is defined by

   $$f(x) = \begin{cases} \dfrac{1}{B-A} & for \quad A < x < B \\ 0 & \text{otherwise} \end{cases}$$

   d. Normal Distribution : The probability for this distribution is determined by calculating the area under the curve of probability density function defined by

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

   e. Chi-Square Distribution : If the random variable $X$ has chi-square distribution with n-degrees of freedom, then the probability density function of $X$ is given by

   $$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1} & \text{if } x > 0 \\ 0 & otherwise \end{cases}$$

   here Γ (.) is a gamma function and it is defined as

   $$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

f. Mathematical expectation or Expected values or Expectations $E(X)$ is defined as $E(X) = x_0p_0+ x_1p_1+ x_2p_2+$…………..$=\Sigma x_ip_i$ when all probabilities are equal then $E(X)$ = mean of $x_0, x_1, x_2$…… Similarly for continuous variables X having density function p(x) where $P[X=x] = p(x)$ , the Expectation $E(X)$ will be given by integral of $x_ip(x_i)$ w.r.t x .
This concept of Expectation also contributes to the definition of Moment Generating Function of *X i.e* $M_x(t)= E(e^{tx})$

## 1.6 SOLUTIONS

**Check Your Progress 1**

**E1** a) If *X* denotes the number of correct answers, then *X* is the random variable for this situation
*b)* *X* can take the values 0,1,2,3…..up to 50
*c)* $P[X=10]$ means the probability that the number of correct answers is 10

**Check Your Progress 2**

**E1** Case (a) is not discrete where as case (b) and (c) are discrete because in case (a) we are taking values in an interval but in case(b) the number of accident is finite , similarly you argue for case (c)

**Check Your Progress 3**

**E1** This situation follows the binomial distribution with n=4 and p=90/100=9/10
The random variable X is the number of seeds that germinate. We have to calculate the probability that exactly two of the four seedswill germinate. That is P[X=2]. By applying Binomial formula , we get

$$P[X=2] = ({}^4C_2) * (9/10)^2 * (1/10)^2$$

$$= 6 * (81/100) * (1/100) = 486/10000 = 0.0486$$

So , the required probability is 0.0486

**Check Your Progress 4**

E1 Here we are dealing with the problem related to the receipt of bad Cheques, which is an event with rare occurrence over an interval of time ( which is a day In this case). So, we can apply Poisson distribution

Average bad Cheques received per day = 6
Thus by substituting $\lambda = 6$ and x=4 in Poisson formula we get

$$P[X=4] = (6^4 e^{-6})/4! = 0.135$$

**Check your progress 5**

E1 Suppose X denotes the time the fire alarm starts measured in minutes after 10:00 am. Then clearly X is a uniform random variable between (0,180). If we want to find the probability that fire alarm will ring before noon, then

$$P(X \leq 12:00 \text{ noon}) = \frac{(12-10)\times 60}{180} = \frac{2}{3}.$$

**Check your progress 6**

E1 a) 0.9332

b) 0.1151

c) 0.9599

E2 The standard normal probability corresponding to this probability is given by

P[952 < Z < 956] = P[((952-952)/4) < ((X-952)/4) < ((952-956)/4)]
= P [0 < Z < 1]
= F(1) - F(0)
= 0.8413 – 0.5 = 0.343

**Check your progress 6**

E1 The moment generating function of $Z = X_1 + X_2$ is

$$M(t) = E\left[e^{t(X_1+X_2)}\right] = E(e^{tX_1})E(e^{tX_2}) = (1-2t)^{-v_1/2}(1-2t)^{-v_2/2} = (1-2t)^{-(v_1+v_2)/2}$$

using **Example 9**.

E2 Using the table in for Chi Square distribution we find in the column headed $x^2_{.95}$ the values: (a) 25.0 corresponding to $v = 15$; (b) 32.7 corresponding to $v = 21$; (c) 67.5 corresponding to $v = 50$.