UNIT 9 PROBABILITY DISTRIBUTIONS

Structure **Page Nos.**

9.1. Introduction 35
Objectives
9.2. Random Variable 36
9.3. Binomial Distribution 42
9.4. Poisson Distribution 46
9.5. Summary 49
9.6. Comments on Exercises 49
9.7. Appendix : Tables of $e^{-\lambda}$ 52

9.1 INTRODUCTION

In the previous unit we had focussed on 'random' and 'random experiment'. We had given several examples to help students understand these concepts. In Sec. 9.2 of this unit, we use these examples to help students relate the outcomes of such experiments to the concept of 'random variable'. We also suggest various real-life situations and activities to help children learn to find the probability distribution of a random variable.

In probability theory there are certain well-known models which describe certain random experiments reasonably accurately. Depending on the nature of values that the random variable takes, they are classified as discrete or continuous models. In Sec. 9.3 and Sec. 9.4 we suggest ways of introducing to students commonly used discrete probability models, the binomial and Poisson models, and their probability distributions. We also consider the way a student can compare the probability distribution of a random variable with these probability distributions, and use the model which has a probability distribution close to that of the random variable.

Not all random variables take discrete values. Some, like the temperature of a place at a particular time, can take any value in an interval. Their probability distributions are continuous. Though we discuss ways of introducing continuous random variables to your learners, we don't dwell on their distributions because the students' syllabus does not include this study.

Throughout the unit we have suggested various exercises and activities for you and your learners to do. Doing them would help the students to get over many of the misconceptions they have regarding the probability distributions they study.

Objectives

After reading this unit, you should be able to develop in your learners the ability to

- explain what a random variable is;
- specify the random variable to be considered in a given situation;
- classify a given random variable as discrete or continuous;
- describe the binomial and Poisson distributions, and calculate the mean and variance associated with these distributions;
- give examples of real-life situations which can be modelled by these discrete distributions.

Following this the students calculate the probabilities as follows:

$P[X = 0] = P\{a_8\} = 1/8$,

$P[X = 1] = P\{a_5, a_6, a_7\} = P\{a_5\} + P\{a_6\} + P\{a_7\} = 3/8$

$P[X = 2] = 3/8$ and $P[X = 3] = 1/8$.

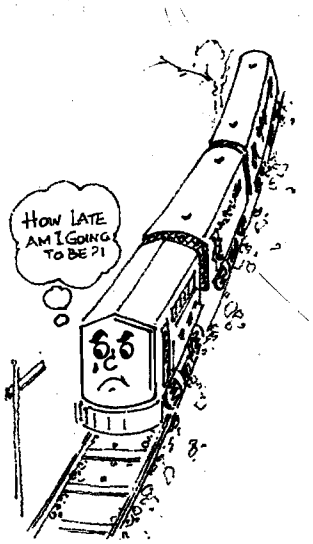
Here is where he insists that they check that P is satisfying the following property:

$P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3] = 1$.

"If they don't add up to 1," he tells them, "you know that you've made a mistake somewhere — either some outcome is left out, or some probability has been wrongly assigned."

Next, Suraj asks the students to break up into groups, and similarly define what the random variable is for some of the situations he gave them to start with — one group discussing the number of phone calls, another group the number of misprints, and so on.

By the next class, Suraj's assessment was that the students had understood that a random variable is a function defined from the sample space of a random experiment, and its range is a set of numbers. By now they had also started using r.v. as an abbreviation for 'random variable'. However, he realised that the examples he had given them were leading them to the misunderstanding that the range of an r.v. is always a subset of $\mathbf{N} \cup \{0\}$. So, he decided to give them the following problem for discussion.



Problem 1 : There is a report in the newspaper that many trains are usually late in our country. Suppose you want to find how late they are. What is the random variable here, its domain and range?

Solution *(done through a peer group discussion, and with Suraj)* : Let us first assume that no train is later than 6 hours, and a few are on time. Now, we want to find the amount of delay for each train. Define the random variable X from the set of trains, such that for each train T_i, $X(T_i)$ is the time (in minutes) by which it is late. So, the variable takes values continuously along a line, say, from time duration 0 to time duration 360 minutes. Any value in between 0 and 360, like 52.33, is possible. $X = 0$ for all trains that are on time, $X = r$ for any r such that $0 < r \leq 360$, for those trains that are r minutes late. In other words, there is no break in the values assumed by this random variable (see Fig. 1).

Fig. 1

So, in this case, the range is an interval, not a subset of **N**.

From more such examples he helped them realise that random variables are of two types : discrete and continuous. He also explained to them the basis on which this categorisation was done and gave them some exercises like the one in E1 to do.

———— × ————

The teacher in Example 1 gave his learners several opportunities to explore the concept of rv in the context of random experiments from their own environment. While thinking about his strategy, you can try the following exercises.

E1) Give your learners the following problem to discuss with each other and do.

Suppose you take a 50-question multiple-choice exam, guessing every answer, and are interested in the number of correct answers obtained. Then

i) What is the random variable you will consider for this situation?

ii) *What values might this random variable take?*
iii) *What would P[X = 40] mean?*

What kind of questions, arguments and misunderstandings of the students showed up in the peer group discussions?

E2) Do you think the strategy given in Example 1 is good? If not, how would you modify it?

Sh. Suraj, in the example above, went on to give his students formal definitions of a discrete rv and its probability mass function (pmf). Alongside he gave many related examples and exercises also.

Regarding these concepts, many students have several misunderstandings and confusions. Some of them are:

- how can a variable be a function?
- how can the rv X be the domain of its pmf? It is not a set.

Unless you talk to your students, and ask them various questions regarding such confusions, they will cope with the course by covering up somehow without understanding what it is about. But, if you want them to build their understanding, you would need to clear their doubts through a variety of examples (like Suraj has done in Example 1).

What many of us need to be clear about ourselves is the definition of a pmf.

Definition : Let S be the sample space of a random experiment and X be the discrete random variable associated with it. We define a function

$p : X(S) \rightarrow [0,1]$ by $p(x_i) = p_i$, where $\sum_{i=0}^{\infty} p_i = 1$ and $X(S) = \{x_i \mid i = 0,1,2,..\}$.

p is called the **probability mass function** (p.m.f.) of the discrete random variable X.

The **graph of the function p**, that is, the collection of pairs (x_i,p_i), i = 0,1,2,... is called **the probability distribution** of X.

For the students who are familiar with frequency distributions, we need to help them see how probability distributions are analogous to them. Asking students to find the pmf of, say, the r.v. denoting the number of heads obtained in three tosses of a coin, is a good idea. Can they find the probability distribution corresponding to this random variable as the set $\left\{\left(0,\frac{1}{8}\right),\left(1,\frac{3}{8}\right),\left(2,\frac{3}{8}\right),\left(3,\frac{1}{8}\right)\right\}$? Can they express this in tabular

Table 1 : Probability distribution of number of heads in three tosses of a coin

Number of Heads (Values of the rv X)	Probability
0	1/8
1	3/8
2	3/8
3	1/8

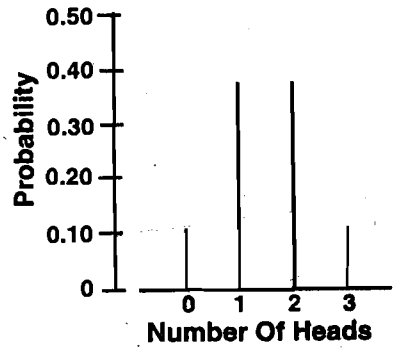


Fig. 2 : Probability distribution of number of heads in three tosses of a coin

form (as in Table 1), and graphical form (as in Fig. 2)? Regarding Table 1, you could point out that it could be thought of as a frequency distribution. The distribution

(shown in the table or graph) tells us how the total probability 'one' is distributed over the possible values of the random variable.

Students who see the probability distribution of a random variable as analogous to a frequency distribution, may ask about the analogue of the mean. It is useful to give real-life examples of the need for the mean of a probability distribution. Consider the following situation.

Problem 2 : The Director of a breast cancer screening clinic wants to make her clinic more efficient. For this she needs to know how many women would be screened on a typical day. The past daily records of the clinic indicate that the number of women screened daily ranges between 100 to 115. Table 2 illustrates the number of times this level, between 100 to 115, has been reached during the past 100 days.

Table 2 : Number of women screened daily during 100 days

Number Screened (x_i)	Number of Days Observed Level (f_i)	$P[X = x_i] = p_i \left(= \frac{f_i}{\Sigma f_i} \right)$
100	1	0.01
101	2	0.02
102	3	0.03
103	5	0.05
104	6	0.06
105	7	0.07
106	9	0.09
107	10	0.10
108	12	0.12
109	11	0.11
110	9	0.09
111	8	0.08
112	6	0.06
113	5	0.05
114	4	0.04
115	2	0.02
Total	100	1.00

On an average how many women per day would this clinic be screening?

Solution : Let us first describe the 'random variable' of interest in this problem. It is the number of patients screened on any given day. The values x_i that this rv can take are given in the 1st column of Table 2. The 2nd column contains the frequency f_i of each value. The last column gives the probability (which is the relative frequency, i.e., $p_i = f_i / \Sigma f_i$ for which a particular value is observed. Notice that the sum of the values in the last column is one. The probability distribution of the rv is graphed in Fig. 3.

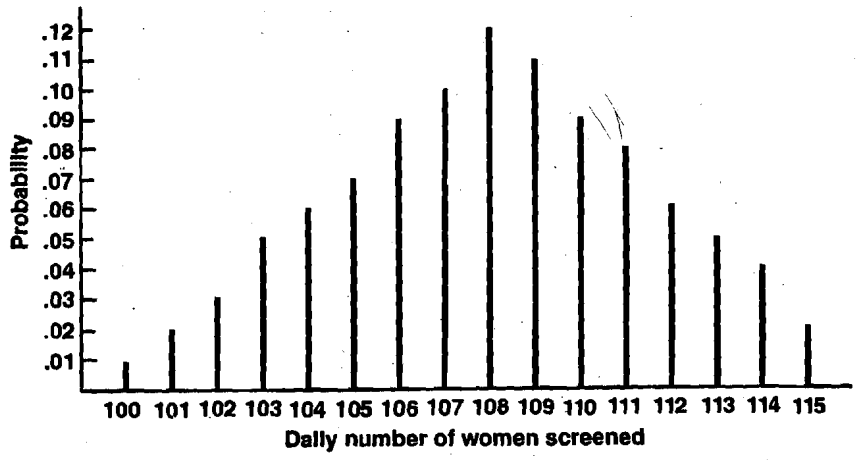


Fig. 3 : Probability distribution for the discrete random variable 'number screened'.

Now, to find the requirement on a typical day, we need to find this average number screened. This is the mean, $\frac{\sum x_i f_i}{\sum f_i} = \sum x_i \left(\frac{f_i}{\sum f_i} \right) = \sum x_i p_i$.

So, the mean, $E(X) = \frac{(100 \times 1) + (101 \times 2) + ... + (115 \times 2)}{100} = 108.02$

This tells us that **over a long period of time**, the number of daily screenings should average about 108. Now, based on this **expected value** (or mean), the director can decide on the resources/infrastructure required for dealing with the expected number of people.

Regarding the students' understanding of the expected value (or mean) of an rv, many of them commonly think that, for instance in the situation of Problem 2, 108 women will visit the clinic every day. You need to stress that $E(X) = 108$ only means that in the **long run** an average of 108 women would visit the clinic. Through several examples, you can help students realise that **the mean is a long run average**.

Your students could also be given a glimpse of the fact that 'expected value' is a fundamental idea in the study of probability distributions. For many years, the concept has been put to considerable practical use in the insurance industry, and by many others like the Director in Problem 2. Giving them exercises like the following one to do will help them see this point.

E3) A second-hand car dealer has sold as many as five cars in one day, and as few as one. She has tabulated sales records for a larger number of days and found that on 5 percent of the days no cars were sold. She took 0.05 as the probability of zero sales in a day, as shown in Table 3 below. Probabilities for sales of 1,2,3,4 and 5 cars were assigned in the same manner (see table below)

Table 3

Number of Cars Sold Per Day	0	1	2	3	4	5
Probability	0.05	0.15	0.35	0.25	0.12	0.08

She wants to find how many cars per day will be sold on the average over a long period. How can she get this number?

So far we have focussed on discrete random variables. Analogous concepts hold good for continuous random variables. The methods we have mentioned above could also be used for helping students understand the concepts related to such rvs. However, you would need to bring to their notice that the major difference is that the possible values that such an rv can take are uncountable. It can take all values in an interval, say,]a, b[. So, we cannot really speak of the i^{th} value of X. Therefore, $p(x_i)$ becomes meaningless in this context. This is why, for defining the distribution of a continuous rv, we replace x_i by any interval of the type $]x_{i-1}, x_i[$, where $a \leq x_{i-1} < x_i \leq b$, and define the probability for such intervals as the area over this interval and under the graph of f, the pdf (see Fig. 4).

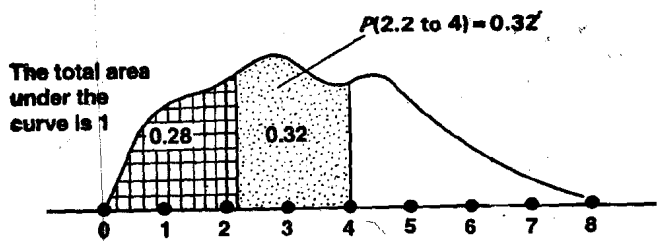


Fig. 4 : Probability distribution of a continuous rv

You should ask your students to 'compare and contrast' the graphs of distributions of discrete and continuous rvs, as for example in Fig. 3 and Fig. 4. This would help them to get some idea of the difference in these types of random variables.

In this context, you would need to think of examples to make the point that the probability that a continuous rv takes a particular value, say x_i, is zero because the area above the point and below the curve is the area of a line segment, which is zero. In general, the **probability that a continuous random variable takes on a particular value is zero**. Consequently, the probability of an interval is the same whether the endpoints are included or not — because the endpoints have probability zero.

Though continuous distributions are not in the syllabus of students, it would be a good idea to expose them to one or two examples of the distribution of a continuous random variable in the context of real-life problems. One such problem is given below.

Problem 3 : Suppose the Director of Personnel in a company wants to conduct a training programme to upgrade the supervisory skills of production line supervisors. Because the programme is self-administered, each supervisor requires a different number of hours to complete the programme. Based on a past study of participants, the following distribution (see Fig. 5) showing the time spent by trainees is available. This shows that the average time spent is 500 hours.

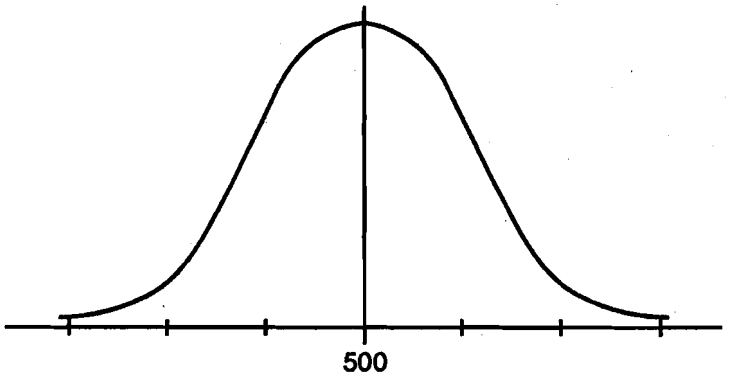


Fig. 5

How can the Director use the graph of the distribution to find the chance that a participant selected at random will require

i) more than 500 hours to complete the programme?
ii) less than 500 hours to complete the programme?

Solution :

i) From the graph of the distribution, we see that half of the area under the curve is located on either side of the mean of 500 hours. So, we conclude that the probability that the random variable will take on a value higher than 500 is ½ , or 0.5.

ii) A similar argument shows that the chance is 0.5 in this case.

Another argument could be : the total probability is 1, (i) and (ii) are mutually exclusive and cover all the possibilities; so, the probability of this event is 1– 0.5 = 0..5.

Of course, exercises like some of the following ones also need to be given to your students. Why don't you try them too?

E4) Which of the random variables given below are discrete? Give reasons for your answer.

i) The daily measurement of snowfall at Shimla.

ii) The number of industrial accidents in each month.

iii) The number of defective goods in a shipment of goods from a manufacturer.

E5) A box contains twice as many red marbles as green marbles. One marble is drawn at random from the box and replaced; then a second marble is drawn at random from the box. If both marbles are green, you win Rs. 50; if both are red, you lose Rs. 10; and if they are of different colours, you will win or lose nothing. What is the probability distribution of the amount you win or lose?

E6) What are the various ways you have used for evaluating how far your students understand 'random variable' and related concepts? What did you conclude from this evaluation?

In the following sections we shall consider ways of introducing students to some distributions in their syllabus.

9.3 BINOMIAL DISTRIBUTION

While introducing your learners to probability, one of the examples you would take is of tossing a coin. I'm sure you usually use this example for introducing them to binomial experiments, that is, experiments where there are only two possible outcomes. You can ask your students to think up many other examples from their surroundings. For example, on tossing a die you either get a four or you don't, a newborn is either a girl or a boy, and so on.

Concrete activities are a good way to get students to grasp a concept. An easy activity for them to do in the class is the experiment of tossing a fair coin several times. They see that each trial of the experiment has **only two possible outcomes** — a head or a tail. You can link these outcome with 'success' and 'failure'. You could, through hints, get your students to note that all the trials are independent of each other, and the probability of getting an outcome H(or T) remains the same in each trial. Of course, they know that the probability of getting a head (or a tail) in a trial is ½. You can ask them what the random variable would be in finding the total number of heads obtained in tossing the coin 3 times. What values would this rv take? If they have done this experiment earlier, they can restate the probabilities in terms of p, the probability of a success (i.e., getting a head) and q, the probability of a failure (i.e., getting a tail). So,

$$P[X = 0] = P\{T,T,T\} = q \times q \times q = q^3 = \frac{1}{8}$$

Similarly, $P[X = 1] = P\{(T,T,H), (T,H,T), (H,T,T)\}$

$$= P\{(T,T,H)\} + P\{(T,H,T)\} + P\{(H,T,T)\}$$

$$= q^2p + q^2p + q^2p = 3q^2p,$$

$P[X = 2] = 3p^2q$, and

$P[X = 3] = p^3$

Again, you could, through hints, help them rewrite the probabilities as

$P[X = 0] = C(3,0)p^0q^{3-0}$

$P[X = 1] = C(3,1)p^1q^{3-1}$

$P[X = 2] = C(3,2)p^2q^{3-2}$

$P[X = 3] = C(3,3)p^3q^0$

You could get them to do a similar exercise for 5 trials, 10 trials, and so on. Each time, they need to observe the following:

i) The trials are independent of each other.
ii) Each trial has two possible outcomes.
iii) The probabilities of a 'success' (p) and of a 'failure' (q) do not change.

Through a variety of examples, you could help them arrive at the generalisation that the probability $P[X = r] = C(n,r)p^rq^{n-r}$, where r = number of successes, n = number of trials made, p = probability of success in a trial, q = 1 − p = probability of failure in a trial.

Fig. 6 : James Bernoulli (1654-1705)

At this stage it would be useful for your learners to sum up the points observed in the activity they have done. They would need to be given the formal definition too.

Somewhere you could mention that these trials are called **Bernoulli trials**, after the seventeenth century Swiss mathematician, James Bernoulli. He did a lot of the early work on binomial distributions.

An activity of the following kind given to the students to do in groups, would be useful in making the distribution more 'real' to them.

Activities *(comparing observed frequencies with those predicted by the binomial distribution) :*

- Ask the different groups to toss 3 dice together 20 times, counting the number of times they get 6/a prime number/a number less than 4/.... They should make a frequency table for the variate they are working on, and compare the observed frequencies with those predicted by the binomial distribution.

- 5% of the population is supposed to be left-handed. Let students find out how many left-handed people are expected in a group of the size of the class. Then they could compare this figure with the actual number found.

You can think of many other activities of this kind. Also, to familiarise your students with real-life situations in which the binomial distribution appears, you could give them problems like the following ones.

Problem 4 : A sales representative calls on four potential clients. The probability that he will obtain an order from each of them is 1/3. Also, whether or not he obtains an order from one of them is statistically independent of whether or not he obtains an order from any of the others. What is the graph of the probability distribution of the number of orders he will receive?

Solution : We note that there are two mutually exclusive events (obtaining an order, or not) each time he makes a call. The probability of an order each time is 1/3. Also the outcomes of the calls are statistically independent. Therefore, this is a situation where there are four Bernoulli trials, and where the probability of a success (getting an order) equals 1/3. So, the probabilities are distributed in the following manner.

$$P[X = 0] = \frac{4!}{0!4!}\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^4 = \frac{16}{81} \approx 0.2$$

$$P[X = 1] = \frac{4!}{1!3!}\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^3 = \frac{32}{81} \approx 0.39$$

'≈' denotes 'is approximately equal to'.

$$P[X = 2] = \frac{4!}{2!2!}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^2 = \frac{24}{81} \approx 0.3$$

$$P[X = 3] = \frac{4!}{3!1!}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^1 = \frac{8}{81} \approx 0.1$$

$$P[X = 4] = \frac{4!}{4!0!}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^0 = \frac{1}{81} \approx 0.01$$

Given the distribution above, we graph it as in Fig. 7.

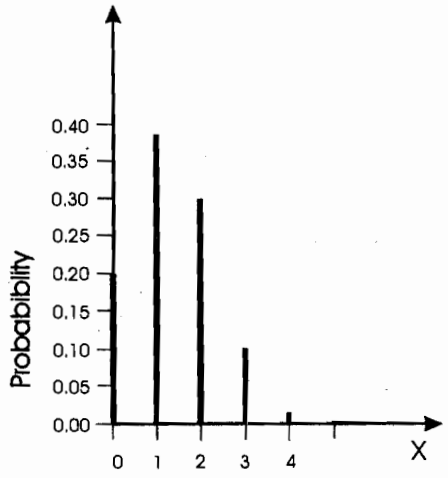


Fig. 7

Problem 5 : It has been claimed that in 60% of all solar heat installations, the utility bill is reduced by at least one-third. What is the probability that the utility bill will be reduced by at least one-third in

i) four of five installations?
ii) at least four of the five installations?

Solution : Here the random variable follows a binomial distribution with p = 0.6. Also, r = 4 and n = 5. To find (i), we have to calculate

$P[X = 4] = C(5,4)(0.6)^4(0.4) = 0.259.$

Now, to find (ii), we have to find the probability that X is at least 4. This probability is the sum of the probabilities that X = 4 and X = 5.

Now, $P[X = 5] = C(5,5)\ (0.6)^5 = 0.078.$

So, the required probability = 0.259 + 0.078 = 0.337.

The binomial distribution is applicable in several other situations, which you could tell your students about. You could give them problems related to deciding about whether to accept a set of goods coming out of a manufacturing process based on how many defective items are in the set. Essentially, the students need to do enough activities so that they recognise common situations in which the binomial distribution is applied, that is, situations **modelled by the binomial distribution**. Along with the examples you expose them to, you should give them exercises like the following ones to do. Why don't you try these ones too?

E7) A farmer buys a quantity of saplings from a company that claims that approximately 80% of the saplings will take root if planted properly. If four

saplings are planted, what is the probability that exactly two will take root?

E8) Consider again the problem of Sunil, the newspaper boy in E5, Unit 8. When a statistics student saw the data collected by him, she started wondering if the number of customers from among Sunil's ten irregular customers, who actually buy from him on a given day, will follow a binomial distribution. Under what conditions would this random variable follow a binomial distribution?

Regarding the mean of a binomial distribution, it may be useful to first present the students with situations like the following ones, and see how they try and solve it.

Problem 6 *: An oil exploration firm plans to drill some holes. It is believed that in the long run, on an average, oil is found in six out of 10 holes dug in this region. Assume that the outcome of drilling one hole is statistically independent of that of drilling any of the other holes. If six holes are dug here, what is the probability of a hole yielding oil.*

Further, the firm will be able to stay in business only if two or more holes produce oil. What is the probability of its staying in business?

It is interesting to observe the students' reactions. Based on openings you get from listening to their arguments and 'solutions', you could hint to them (if necessary) that the 'mean' may be required here.

Your students would already know that for a discrete random variable X, the expected value, $E(X) = x_0p_0 + x_1p_1 + \ldots$, where $x_0, x_1, \ldots$ are the values assumed by X, and $P[X = x_i] = p_i \ \forall \ i = 0,1,2,\ldots$

So, if X is a binomial rv, taking values 0,1,…, n, they would know that
$E(X) = n \times C(n,n)p^n(1-p)^0 + (n-1)C(n,n-1)p^{n-1}(1-p)^1 + \ldots + 0 \times C(n,0)p^0(1-p)^n$
You can give them a hint about manipulating the RHS to get

$$E(X) = np \sum_{j=1}^{n-1} C(n-1, j-1)p^{j}(1-p)^{n-j}$$

$$= np[1-p+p]^{n-1}$$

$$= np.$$

This means that the **expected number** of successes is np.

Having found this, they would be in a position to solve Problem 6, may be as given below.

Solution to Problem 6 : To start with, each hole drilled can be viewed as a Bernoulli trial where the probability of success is p. We know that the mean is 0.6.
So, $np = 0.6 \Rightarrow \ 6 \times p = 0.6 \Rightarrow p = 0.1$.

Now, the probability that the firm stays in business is
[1 − (prob. of getting 0 or 1 oil-producing holes)]

$$= 1 - P(X = 0 \text{ or } 1) = 1 - \left[\frac{6!}{0!6!}(.9^6) + \frac{6!}{1!5!}(.1)(.9)^5\right] = 0.115.$$

Why don't you try some exercises now?

E9) List 5 problems on the mean and variance of the binomial distribution, each related to a real-life situation from your learners' environment.

E10) What are the limitations of the binomial distribution? How would you help your learners become aware of them?

In the next section we suggest ways of introducing your learners to another discrete distribution, named after the 19th century French mathematician, Poisson.

9.4 POISSON DISTRIBUTION

In the standard texts available to your learners at present, the Poisson distribution is presented as a limiting case of the binomial distribution. But the examples are not such that help them develop an understanding about which of these distributions should be used when. For this you could ask them to consider the following situation, for example.

Suppose it is the busy Friday noon hour at a bank, 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 given time interval**. Let us make some assumptions.

Fig.8 : Simeon-Denis Poisson (1781-1840)

1) The average arrival rate at any unit time remains the same over the entire noon hour.

2) The number of arrivals in a time interval does not depend on what happened in previous time intervals.

3) The time interval is divided into n sub-intervals. In each of these, it is extremely unlikely that there will be more than one arrival. This means that it is impossible for more than one customer to get through the revolving entrance door in the short unit time interval.

So, in any of these short intervals, either 0 or 1 person can come. So, the binomial distribution can be used over the n intervals that make up the noon hour, where n is very large. Now, if the probability of persons arriving in the short interval is very small, say 0.05, then calculating $P[X = r] = C(n,r)(0.05)^r (0.95)^{n-r}$ is very tedious for large n. Here is where the Poisson distribution is useful. For this we make use of the **Poisson formula**, given by

$$P[X = r] = \frac{e^{-\lambda t}(\lambda t)^r}{r!}, r = 0,1,...$$

where λ denotes the average arrival rate per unit of time and r is the number of arrivals in t units of time.

Suppose we know that $\lambda = 72$ arrivals per hour is a constant for this situation, and we want to find the probability of 4 arrivals in 3 minutes. Since λ is given in 'hours', we first need to standardise the unit and find 't' in hours. So t = 3 min. = $\frac{1}{20}$ hours.

Then $P[X = 4] = \frac{e^{-\frac{72}{20}}\left(72 \times \frac{1}{20}\right)^4}{4!} = \frac{e^{-3.6}(3.6)^4}{4!} = 0.191$, using the table given in the appendix to this unit.

What does this value 0.191 tell us? It says that there is a 19.1% chance that exactly four customers will arrive in the next 3 minutes.

If we vary the values of r and t, we can get different probabilities. This gives the probability distribution, which is called the Poisson probability distribution. The graphs for a few different values of λ are given in Fig. 9 below.

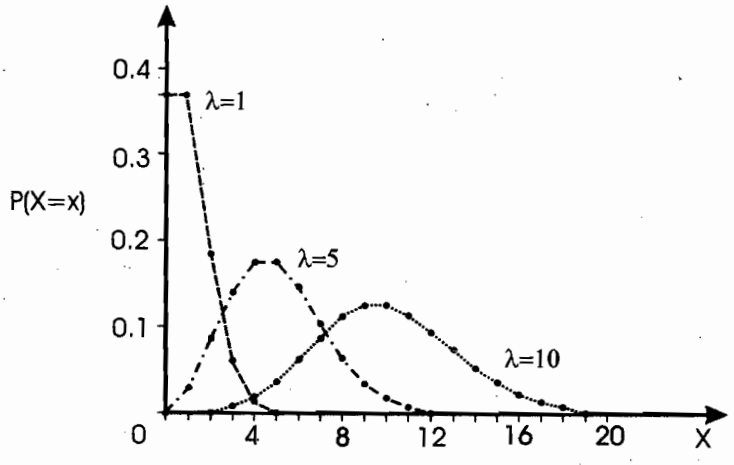


Fig. 9

A useful activity to help learners see the utility of the Poisson distribution is the 'Traffic-flow study'.

Activity : *To investigate if the number of people entering a particular public area in a given time interval follows a Poisson distribution.*

You can divide your students into groups. Ask each group to observe the number of people entering the canteen (or the bathroom, or the library, etc.) in a given time interval, for 50 or more consecutive intervals. This activity should be done at a reasonably busy time for the area chosen. The time interval should be chosen so that on an average about 3 or 4 people enter during it. Ask the students to make a frequency table with the number of people as the variate. Ask them to calculate the mean and variance. (For a Poisson distribution they should be approximately equal.)

You could also give your learners problems involving a situation where this distribution is applied, like the following one.

Problem 7 : The calls at a telephone switchboard occur at an average rate of six calls per 10 minutes. Suppose the operator leaves for a 5-minute coffee break. What is the probability that exactly two calls come in while he is away?

Solution : The students should **check that the conditions of the Poisson formula are satisfied** in this case. Here $\lambda = \frac{6}{10}$ and t = 5, so that $\lambda t = 3$.

Hence, the required probability P[X = 2] is $\frac{e^{-3}3^2}{2!} = 0.224$.

This means that there is a 22.4% chance that two calls go unanswered in this period.

Here are some exercises for you and your learners now.

E11) If a bank receives 6 bad checks per day on an average, what is the probability that it will receive 4 bad checks on any given day?

E12) A hospital has 20 kidney dialysis machines and the chance of any one of them malfunctioning on any day is .02. To find the probability that exactly 3 machines will be out of service on the same day, use the

i) binomial formula.
ii) Poisson formula.

The point of doing E12 is that your students realise that the difference between the two calculations is very small. Either the binomial or Poisson formula can be applied. The convention followed by many statisticians is that if $n \geq 20$ and $p \leq 0.05$, then the Poisson formula can be used to calculate the probability. It is also clear that the Poisson calculation is simpler than the binomial calculation since it involves only one parameter, λ. Consequently, the Poisson probabilities can be tabulated more compactly than the binomial probabilities. For example, the Poisson probability $P[X = 3]$ is the same for $n = 200$, $p = 0.01$ as it is for $n = 100$, $p = 0.02$, and for any other pair of n and p values whose product is $\lambda = np = 2$.

By now your learners would have got a fairly good idea where the Poisson formula can be used. However, there is an example of its use that they may not have come across. Here the probability calculated is not over an interval of time, but over a region (or space) or something else as our physical reference.

During a war, a rocket hit South London. Later a study was conducted on which regions were not affected by the rocket hit. For this study, λ, the average number of hits per unit area was first calculated. (Note that usually in the Poisson formula, λ is the average rate per unit time.) To calculate λ, the researchers divided the area into 576 areas of equal size (the number 576 was chosen based on some other study), and they found that there were 537 hits.

So, the average number of hits per unit area, $\lambda = \frac{537}{576} = 0.9323.$

The researchers assumed that all the conditions to satisfy the Poisson formula, $P(r) = \frac{e^{-\lambda v}(\lambda v)^r}{r!}$, *hold in this case. According to the problem stated, they had to calculate the probability of 'no hit' per unit area, that is, $P[X = 0]$ and $v = 1$, v being the number of units of the unit area. So, the required probability was*

$P[X=0] = e^{-0.9323} = 0.3936.$

Hence, out of 576 regions, the number of regions not hit by the rocket would be $576 \times 0.3936 = 226$. In fact, the actual number got from the record was 229 regions, quite close to the value got by using the Poisson formula.

Such problems help your students see that **the Poisson distribution is very effective in studying various real-life problems involving a discrete rv, where the occurrence is rare.**

Why don't you try an exercise now?

E13) List 2 real-life situations which your students would relate to and which are effectively modelled by the Poisson distribution.

E14) How is the Poisson distribution used in biological research for counting the number of cells of a particular type in a dilute solution?

One of the main disadvantages of the Poisson distribution is that it is applicable only in situations where the outcomes are independent, i.e., each outcome is independent of what happened previously. There are other distributions that overcome this shortcoming, but are beyond the purview of this course.

With this we come to the end of this unit.

Let us now summarise the points we have covered in this unit.

9.5 SUMMARY

In this unit we have discussed ways of communicating the following points to your learners.

1) A random variable is a function from the sample space of the experiment concerned to a set of numbers.

2) There are two types of random variables — discrete and continuous. (We have mostly discussed aspects of discrete random variables in this unit.)

3) A probability distribution of an rv gives us the probability shared out among the various values in the range of the rv. A pictorial representation is very useful for grasping this sharing.

4) The expected value of a discrete rv X is $E(X) = \sum x_i p_i$, where the range of X is $\{x_0, x_1, \ldots\}$ and $p_i = P[X = x_i]$ $\forall i = 0,1,...$

5) **The binomial distribution** : This is applicable for modelling a series of independent trials, each trial having only two outcomes, 'success' and 'failure'. In n independent trials of the experiment, the probability of an event P[X=r] in this distribution is given by

$P[X = r] = C(n,r)p^r q^{n-r}$, $0 \leq r \leq n$,

where p is the probability of getting a success, $q = 1-p$, and p and q remain the same in each trial.

The mean of the distribution is $E(X) = np$.

6) **The Poisson distribution** : This is applicable in the same situations in which the binomial distribution is applicable. However, it is useful to apply when n is very large and p is very small. The probability of an event P[X = x] in this distribution is given by

$P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$, where λ is the mean (and variance) of the distribution, and is a constant in a particular situation.

9.6 COMMENTS ON EXERCISES

E1) i) If X denotes the number of correct answers, then X is the random variable for this situation.
ii) X can take any of the values 0,1,2..., 50.
iii) The probability that the number of correct answers is 40.

Our learners' thought processes show up in their discussions. These give us a clue to how they grasp concepts and link them with other concepts. Therefore, it is important to share in this aspect.

E2) Did you try this method with your students? What pros and cons did you find? What teaching-learning points showed up in the process? How much more time did you take out to evaluate the level of understanding of your learners, or was that built into the strategy? What other aspects have you gone into regarding Suraj's strategy?

E3) She has to calculate the mean. It is given by

$$\text{Mean} = 0 \times 0.5 + 1 \times 0.15 + 2 \times 0.35 + 3 \times 0.25 + 4 \times 0.12 + 5 \times 0.08$$
$$= 2.48$$

This means that she can expect that on an average 3 cars will be sold per day in the long run.

E4) The rv in (i) is not discrete because it can take all values in an interval. The rvs in (ii) and (iii) are discrete because the number of accidents, as well as of defective goods, is finite.

E5) Let X denote the amount you win or lose. Then X takes values Rs. 50, 0 or −10 (loss of Rs. 10). Suppose there are n green and 2n red marbles. The probability that both the marbles are green is $\frac{n}{3n} \times \frac{n}{3n}$, i.e., P[X=50] = 1/9.

The probability that both the marbles are red is $\frac{2n}{3n} \times \frac{2n}{3n}$ i.e., $P[X=-10]=4/9$.

The probability that the marbles are of different colours is 4/9, i.e., P[X = 0] = 4/9.

Thus the probability distribution is as given in the following table.

Amount	Probability
50	1/9
0	4/9
−10	4/9

E6) Did you use oral and written tests? If so, what kinds of questions did you ask — those that assessed fact retention, or assessed the objectives listed in Sec. 9.1? Did you give them assessment activities to do? If so, of what kind?

How did you go about analysing the results of your evaluation? What did the analysis tell you about your teaching strategy?

E7) This situation follows the binomial distribution with n = 4 and $p = \frac{80}{100} = \frac{4}{5}$.

The random variable X is the number of saplings that take root. We have to calculate the probability that exactly two will take root. This is

$$P[X = 2] = C(4,2) \left(\frac{4}{5}\right)^2 \times \left(\frac{1}{5}\right)^2 = 0.154.$$

E8) If X_i denotes the random variable that the ith customer buys the paper on a given day, then the X_is may not be identically distributed. Therefore, the binomial distribution is not appropriate, unless the customers have the same business activities, or habits, or working nature. Under these circumstances, we can expect the X_is to be identically distributed. And then we can expect the X_is to follow the binomial distribution.

E9) For instance, if the mean and variance of the life of a brand of 'chappals' is known, calculating the probability that a 'chappal' lasts for that period.

You can think of several other situations.

E10) You could give them some problems that help them realise that, for instance, one limitation is that if the number of trials 'n' is very large and probability 'p' is very small, the computation of P[X = r] becomes cumbersome.

E11) The problem deals with the receipt of bad cheques, which is an event with a rare occurrence over an interval of time (a day, in this case). So, we can apply

the Poisson distribution. On an average 6 bad cheques are received per day. So, substituting $\lambda = 6$, $x = 4$ and $t = 1$in the Poisson formula, we get

$$P[X = 4] = \frac{6^4 e^{-6}}{4!} = \frac{1296 \times (0.0025)}{24} = 0.135.$$

E12) Note that here the experiment or trial is 'checking the machine for its functioning'. There are 20 independent trials, and each trial is identically distributed with probability 0.02.

i) Applying the binomial formula, we get

$$P[X = 3] = \frac{20!}{3! \times 17!}(0.02)^3(0.98)^{17}$$

$$= 0.0065$$

ii) Note that here the average rate of machines that go out of service in a day is the constant $\lambda = 20 \times 0.02 = 0.4$.

Also note that we can make the sub-intervals so small that at best only one machine go out of service. Thus, conditions applying the Poisson formula are satisfied. So,

$$P[X = 3] = \frac{(0.4)^3 e^{-0.4}}{3!} = \frac{(0.064)(.67032)}{6} = .00715$$

E13) For your science students, one situation is suggested in E14.
For another situation, you can build a problem supposing that the probability of having no electricity is rare, with an average of 2 days in a semester.

E14) Are the cells randomly distributed in a given solution? To find out, well-shaken solution is placed on a slide which is divided into squares and viewed through a microscope. The number of cells in each of the squares is counted, and from the mean, the number of cells per unit volume can be estimated. Agreement of the observed frequencies with a Poisson distribution with the same mean is used to test that the cells are distributed randomly through the solution.

Alternatively, if it is known that the cells are randomly distributed, a quick method is available for estimating the total number of cells present. For instance, suppose 22 out of the 500 squares on a slide contain no cells.

The relative frequency for the number of squares containing no cells is $\frac{22}{500}$.

Equating this with $P[X = 0]$ gives $\frac{22}{500} = e^{-\lambda}$,

where λ is the mean number of cells per square. Using tables we find $\lambda = 3.124$

So, we estimate the total number of cells present by taking the mean number of cells per square multiplied by the number of squares, i.e. $3.124 \times 500 = 1562$.

APPENDIX

Table 4 : Values of $e^{-\lambda}$, $\lambda > 0$

λ	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	.9048	.8958	.8869	.8781	.8694	.8607	.8521	.8437	.8353	.8270
0.2	.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	.5488	.5434	.5379	.5326	.5273	.5220	.5169	.5117	.5066	.5016
0.7	.4966	.4916	.4868	.4819	.4771	.4724	.4677	.4630	.4584	.4538
0.8	.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	.4066	.4025	.3985	.3946	.3906	.3867	.3829	.3791	.3753	.3716

Table 5 : Values for $\lambda = 1,2,3,\ldots 10$

λ	1	2	3	4	5	6	7	8	9	10
$e^{-\lambda}$	.36788	.13534	.04979	.01832	.00638	.002479	.00091	.000335	.000153	.000045

Note : To obtain values of $e^{-\lambda}$ for other values of λ, use the laws of exponents.

For example, $e^{-2.35} = (e^{-2.00})\,(e^{-0.35}) = (.13543)\,(.7047) = .095374$.