SUMMARY

Probability for Statistics

In this module, you will learn about inferential statistics and understand how it can save a lot of time and effort while working with large sets of data. Exploratory data analysis (EDA) helps you discover patterns in data using various techniques and approaches; however, it is also the part on which data analysts spend most of their time. On the other hand, inferential statistics enables you to 'infer' insights from sample data.

Introduction to Inferential Statistics

Inferential statistics is a widely used technique to deal with large datasets. As previously mentioned, inferential statistics is a method of inferring insights from a sample set of data. It enables one to first gather inferences based on a small subset of the given data and then analyse it.

Descriptive statistics, on the other hand, is a method of inferring insights from the entire population; however, this might not be possible in every situation. This is where inferential statistics comes into play, as it allows one to use a sample of the provided data and make inferences from the same. Since you will be working with a sample of the entire data, the results will not represent the exact values; instead, they will be in terms of probability.

Introduction to Probability

This segment helped you understand probability in detail and the various approaches that are used to calculate the probability of an event. It is important to understand how statistics and probability are related to each other in spite of them being two extremely different measures of analysis for a given data set. Let’s recall the differences between the two.

Statistics	**Probability**
It deals with a sample of data.	It deals with a model of random data.
It infers information about a population or a random process that produced a sample.	It deduces information about random events or samples produced from a model.

One of the common examples of probability in our day-to-day lives is the weather report for the day. Let’s say that there is a 40% chance that your city will experience rainfall. This indicates that the probability of rainfall is 0.4. Probability also affects your day-to-day decision-making in the sense that if the probability of rainfall is high, then you may choose to carry an umbrella with you.

Random Variable

A random variable is a variable whose values are the outcomes derived from a random experiment. Let’s understand this using a simple example of a weather forecast that we cited above. Given that the probability of rainfall is 0.4, the outcome of this experiment is ‘it will rain’.

Suppose X is a random variable and the chances of rainfall are 40%. Therefore,
P (X = ‘it will rain’) = 0.4

However, there is a 60% chance of no rainfall.
P (X = ‘it will not rain’) = 0.6.

Another important feature of a random variable X is that it converts the outcomes of experiments to measurable values. Suppose you are a data analyst at a bank. You are trying to find out which customers will default on their loans, i.e., stop paying their loans. Based on some data, you have been able to make predictions that are listed in the table given below.

Customer Number	Yearly Income (in ₹)	Amount of Loan Due (in ₹)	Number of Dependants	Default Prediction (Yes/No)
1	10 lakh	75 lakh	3	Yes
2	15 lakh	50 lakh	2	No
3	20 lakh	40 lakh	1	No

Now, instead of processing the yes/no response, it will be much easier for you to define a random variable X to indicate whether or not the customer is predicted to default. The values will be assigned according to the following rule:

X = 1 if the customer defaults; and

X = 0 if the customer does not default.

Now, the data changes as given below.

Customer Number	Yearly Income (in ₹)	Amount of Loan Due (in ₹)	Number of Dependants	X (Random Variable)
1	10 lakh	75 lakh	3	1
2	15 lakh	50 lakh	2	0
3	20 lakh	40 lakh	1	0

Random Variables and Probability Tables

In this segment, you went through the red ball example. Let’s quickly revise this example. Suppose you go to your friend's place for a party. The host has organised various games for fun and entertainment. You decide to play one of these games.

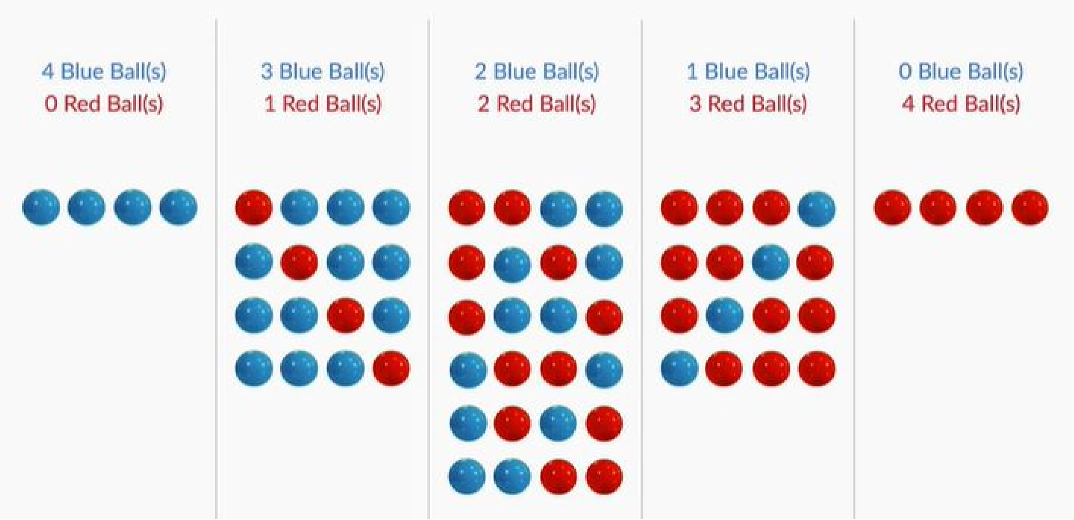
Problem statement: The game requires you to draw a red ball from a bag for four consecutive times. The bag consists of five balls, out of which three balls are red and two balls are blue. Each time you draw a ball, note its colour and put it back in the bag. The entry fee for the game is $1. If you are able to draw the red ball from the bag for four consecutive times, then you win $15.

Question: In the long run (i.e., if you play the game a lot of times), is this game profitable for the players or for the house (host)? Or, will everybody break even in the long run?

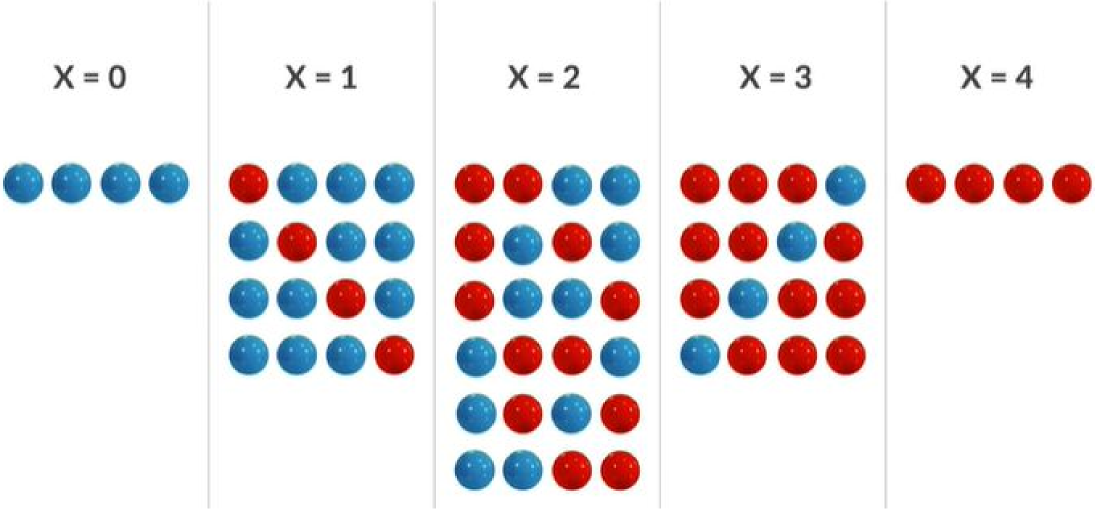
Solution: The answer to this question lies in the following three steps:

1. Find all the possible combinations.
2. Find the probability of each combination.
3. Use the probabilities to estimate the profit/loss per player.

Step 1 involves finding all the possible combinations. Take a look at the image given below that shows all the possible outcomes.



Step 2 involves finding the probability of each combination. Let's first define a random variable X, where X is the number of red balls picked. Take a look at the image given below that shows the value of the random variable X for each combination.



Now, you need to find the probability of all the possible outcomes shown in this image. This can be done using the following formula for probability:

Probability=$Number\ of\ favourable\ outcomes \div Total\ number\ of\ outcomes$

The probability of each random variable is given in the table provided below.

X	P(X)
X = 0	2/75 = 0.027
X = 1	12/75 = 0.160
X = 2	26/75 = 0.347
X = 3	25/75 = 0.333
X = 4	10/75 = 0.133

In this table, you can see that the probability of obtaining the random variable X = 4 is only 0.133, which is extremely low. This indicates that the chances of winning this game are extremely low. Here, most of the players are losing their $1 rather than winning $15.

Probability Distribution

A probability distribution is a form of representation that tells us the probability for all the possible values of X. In a valid and complete probability distribution, there are no negative values, and all the probability values add up to 1. It could be:

- A table,
- A chart, or
- An equation.

In this segment, you also learnt the difference between direct computation and simulation. Let’s recall these concepts. Both direct computation and simulation are essential for calculating probabilities. Although direct computation is accurate, it requires both effort and time and is, therefore, extremely difficult when working with huge data sets. On the other hand, simulation can be performed for large and complex data sets, and it also saves human effort using more computer time, but the results are not as accurate as direct computation. Hence, the choice of method depends on the type of data available.

Expected Value

The expected value (EV) of a variable is the weighted average of all possible values of a random variable X. It enables you to compare the given values and analyse which of the given values is better. In other words, the expected value for a random variable X is the value of X that you would 'expect' to get after performing the experiment an infinite number of times. It is also called the expectation, average or mean value.

Mathematically speaking, for a random variable X that can take the values x1,x2,.....,xn, the expected value (EV) is given by the following formula:

$$EV(X) = x1*P(X = x1) + x2*P(X = x2) + x3*P(X = x3) + ... + xn*P(X = xn)$$

Now, let's reconsider the example of the red ball game. We are yet to determine whether the game was profitable for the players or for the house (host). For doing so, you need to follow the three steps mentioned below:

1. Find all the possible combinations.
2. Find the probability of each combination.
3. Use the probabilities to estimate the profit/loss per player.

The first two steps are already completed. Let’s move on to Step 3, where we will use the probabilities that we calculated to estimate the profit/loss per player. Now, suppose the number of people playing the game was 1,000 instead of 75.

Total number of players = 1000

P(X = 0) = 0.027; number of people that draw 0 red balls = 0.027 * 1000 = 27

P(X = 1) = 0.16; number of people that draw 1 red balls = 0.16 * 1000 = 160

P(X = 2) = 0.347; number of people that draw 2 red balls = 0.347 * 1000 = 347

P(X = 3) = 0.333; number of people that draw 3 red balls = 0.333 * 1000 = 333

P(X=4) = 0.133, Number of people that draw 4 red balls = 0.133 * 1000 = 133

Total number of red balls drawn = 0 ∗ 27 + 1 ∗ 160 + 2 ∗ 347 + 3 ∗ 333 + 4 ∗ 133 = 2385

The average number of red balls per game = 2385/1000 = 2.385

The average number of red balls drawn in one game is the expected value of red balls to be drawn. You may be wondering how you can get 2.385 balls in one attempt. This decimal value is made possible by the fact that the expected value gives the average of all the values obtained if the game is played an infinite number of times.

What does this mean? How does this help us with our original question, which was, 'On average, how much money are the players expected to make?'. In order to figure this out, let's take another random variable. Now, our random variable X signifies the money won by a player after playing a single game.

Here, the random variable X can take only two values, i.e., either -1 (if the player loses) or 15 (if the player wins). The probabilities of X can be defined as follows:

- P(X = 15) = P(4 red balls) = 0.133
- P(X = -1) = P(0 red balls) + P(1 red ball) + P(2 red balls) + P(3 red balls)
    - P(X=-1) = 0.027 + 0.16 + 0.347 + 0.333 = 0.867

Expected value = E[X] = (15 * 0.133) + (-1 * 0.867) = 1.128

This means that on average, a player will win $1.128, and the host will end up with a huge loss. For the host to gain profit from this game, the expected value should come out to be negative. This can be done by decreasing the price money, increasing the entry ticket (i.e., the penalty) or decreasing the chances of winning. This is how casinos earn from their customers.

Next, you also learnt about some properties of the expected value, which are as follows:

- Given the random variable X, E[aX + b] = a E[X] + b
- E[constant] = constant
- Given two independent random variables X and Y, E[X + Y] = E[X] + E[Y]

Standard Deviation of a Random Variable

In this segment, you learnt that the expected value of a variable may fail your analysis, as it works the best for large-sized data. The value can be affected by the presence of outliers; therefore, you need to have a measure that analyses the risk involved in the analysis. This measure is the standard deviation of the variable. The standard deviation of a random variable is simply the standard deviation of a sample as it goes to infinity. It is denoted by 'σ'.

Some additional points related to the standard deviation of the variable are as follows:

- Variance = Var(X) = $E[(X - E[X])]^2 = \Sigma(X - \mu)2p(X)$
- In other words, the variance is the expected value of the squared difference of the random variable X and the expected value m.
- Standard deviation (σ) = $\sqrt{\ }$Var(X)

SUMMARY

Discrete Probability Distributions

In this session, you learnt about discrete probability distributions, i.e., probability distributions that are commonly used for discrete random variables, such as binomial probability distribution and uniform probability distribution.

Probability Without Experiment

In this segment, you learnt how to calculate probabilities without having to perform a lengthy experiment. Let's recall the experiment of drawing red balls. It is very difficult to perform the experiment 75 times or more in order to calculate probabilities. Therefore, you learnt an effective method in order to solve this problem. Let's quickly recall the example.

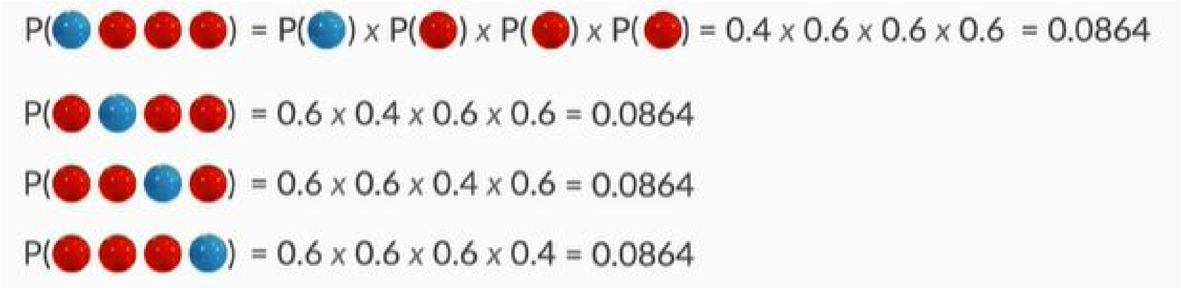
The total number of balls in the bag are 5, out of which 3 are red and 2 are blue. Now, let's try to calculate the probability:

P(1 red ball in 1 trial) = Number of red balls / Total number of balls = 3/5 = 0.6

Multiplication rule: P(Event 1 AND Event 2) = P(Event 1)*P(Event 2). Therefore, we can say that:

P(2 red balls in two trails) = P(1 red ball in the first trial) * P(1 red ball in the second trial) = 0.6 * 0.6 = 0.36

Now, let's understand how the probability of getting 3 red balls is calculated. There can be various combinations of obtaining 3 red balls. Let's first calculate the probability of each of these combinations as shown in the image given below.


P( ) = P( ) x P( ) x P( ) x P( ) = 0.4 x 0.6 x 0.6 x 0.6 = 0.0864
P( ) = 0.6 x 0.4 x 0.6 x 0.6 = 0.0864
P( ) = 0.6 x 0.6 x 0.4 x 0.6 = 0.0864
P( ) = 0.6 x 0.6 x 0.6 x 0.4 = 0.0864

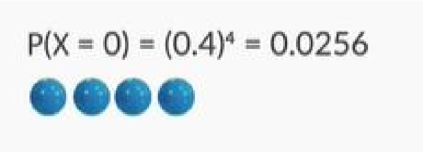
Now, the probability of attaining 3 red balls in total is the sum of all the aforementioned combinations.

Addition rule: P(Event 1 OR Event 2) = P(Event 1) + P(Event2)

P(3 red balls) = P(X=3) = 0.0864 + 0.0864 + 0.0864 + 0.0864 = 4 * 0.0864 = 0.3456

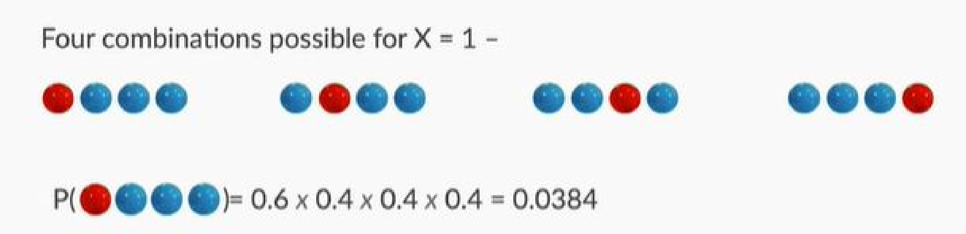
Similarly,

1. The possible combinations for 0 red balls are shown below.



$P(X = 0) = (0.4)^4 = 0.0256$

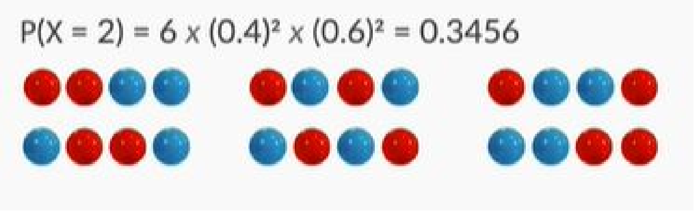
2. The possible combinations for 1 red ball are shown below.



Four combinations possible for X = 1 –

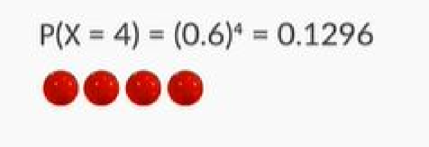
P(🔴🔵🔵🔵)= 0.6 x 0.4 x 0.4 x 0.4 = 0.0384

3. The possible combinations for 2 red balls are shown below.



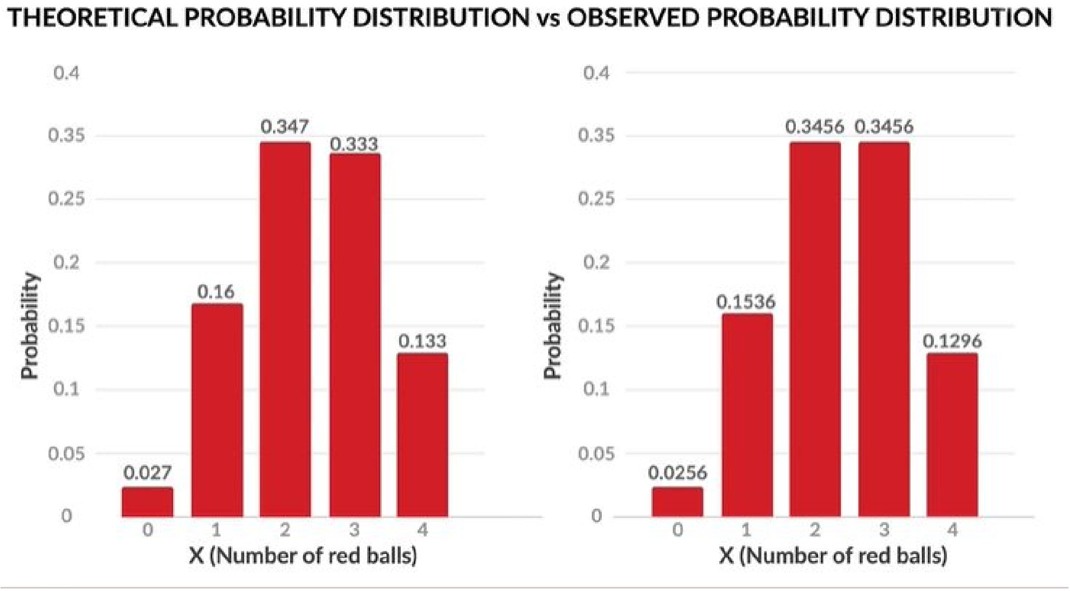
$P(X = 2) = 6 \times (0.4)^2 \times (0.6)^2 = 0.3456$

4. The possible combinations for 4 red balls are shown below.



$P(X = 4) = (0.6)^4 = 0.1296$

Now, let's compare the results of the experiment conducted with 75 people with the results of probabilities computed above by means of a graph as shown below.



As you can see, the graphs shown above have a similar trend, but the values are not exactly the same. The difference between the theoretical and observed values is due to the fact that in the observed probability distribution, you are trying to approximate the results without having to perform the experiments. Therefore, the results are different.

Binomial Distribution

When you flip a coin, there can be two possible outcomes: heads or tails. This makes it a binary random variable. A binary variable is one with two possible outcomes. In this segment, you learnt how to generalise the ball drawing example. Let's revise it.

Suppose the probability of getting 1 red ball in one trial is equal to 'p'. Therefore, the probability of getting 4 red balls, using the multiplication rule, can be given by the following:

$$P(X = 4) = p * p * p * p = p^4$$

Similarly,

$$P(X = 0) = (1 - p) * (1 - p) * (1 - p) * (1 - p) = (1 - p)^4$$

The number of combinations for P(X = 4) and P(X = 0) is 1 each, and hence, by simply applying the multiplication rule, you can obtain the results. On the other hand, in the cases of P(X = 1), P(X = 2) and P(X = 3), you will obtain more than one combination, and hence, you can apply the addition rule to add up all the probabilities to obtain the final result. The calculation for P(X = 1) is given below. As you observed earlier, the number of possible combinations for obtaining 1 red ball is 4. When you add up the results of all the 4 combinations, you will obtain the following result:

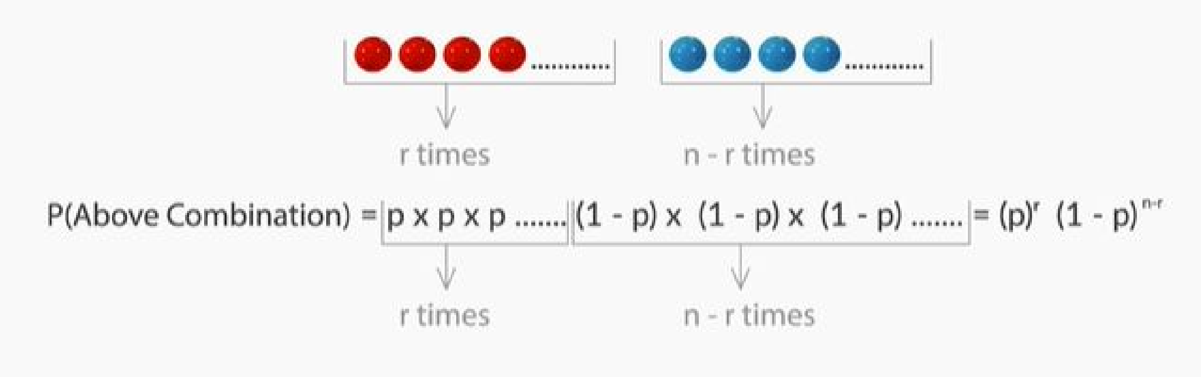
$$P(X = 1) = 4 * p * (1 - p) * (1 - p) * (1 - p) = 4p(1 - p)^3$$

Similarly,

$$P(X = 2) = 6p^2(1 - p)^2$$

$$P(X=3) = 4p^3(1 - p)$$

The example that you saw above is for an event/experiment conducted by one person. However, while working in the industry, an event conducted by one person might not be sufficient for the analysis. Now, let's try to understand this example considering multiple events and use a more generic case. Suppose you need to draw r red balls from the bag, given the probability of drawing a red ball is p. Take a look at the image given below.

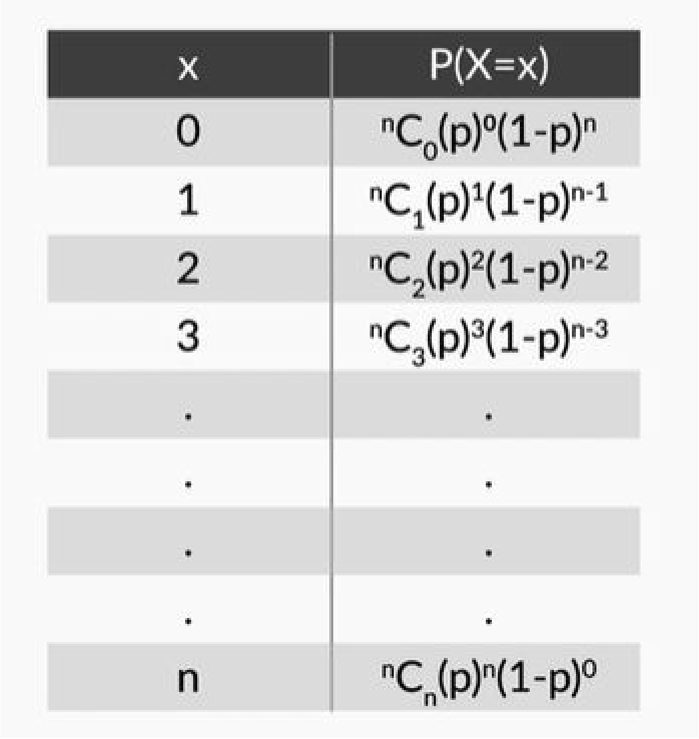

r times
n - r times
P(Above Combination) = p x p x p .... (1 - p) x (1 - p) x (1 - p) .... = (p)^r (1 - p)^n-r
r times
n - r times

The formula for finding binomial probability is given below:

$$P(X = r) = {}^{n}Cr(p)^{r}(1 - p)^{n-r}$$

Where n is the number of trials, p is the probability of success, and r is the number of successes after n trials.

On applying the formula, you will obtain the results shown below.



x	P(X=x)
0	${}^{n}C_{0}(p)^{0}(1-p)^{n}$
1	${}^{n}C_{1}(p)^{1}(1-p)^{n-1}$
2	${}^{n}C_{2}(p)^{2}(1-p)^{n-2}$
3	${}^{n}C_{3}(p)^{3}(1-p)^{n-3}$
.	.
.	.
.	.
.	.
n	${}^{n}C_{n}(p)^{n}(1-p)^{0}$

However, there are some conditions that need to be met in order for us to be able to apply the formula. They are as follows:

1. The total number of trials is fixed at n.
2. Each trial is binary, i.e., it has only two possible outcomes: success or failure.
3. The probability of success is the same in all trials and is denoted by p.

A few examples to understand these conditions in detail are listed in the following table.

Binomial Distribution (Applicable)	Binomial Distribution (Not Applicable)
Tossing a coin 20 times to determine how many tails occur	Tossing a coin until a heads occurs
Asking 200 randomly selected people whether or not they are older than 21 years	Asking 200 randomly selected people how old they are
Drawing 4 red balls from a bag and putting each ball back after drawing it	Drawing 4 red balls from a bag and not putting each ball back after drawing it

So far, you have learnt what binary and binomial variables are and computed their probabilities. The other important measure is the mean and standard deviation of these variables. Now, suppose the probability of the required event is equal to p and the number of times of occurrence is n. Given the probability of the event, the mean and standard deviation for the binary and binomial random variables are as follows:

- Binary variables:
  - Mean = E[X] = p
  - Standard deviation = $\sigma = \sqrt{p(1-p)}$
- Binomial variables:
  - Mean = $n * p$
  - Standard deviation = $\sigma = \sqrt{n * p(1-p)}$

Cumulative Probability

In this segment, you learnt about cumulative probability, which forms the basis of continuous probability, which you will learn about in the next session. Until now, you only calculated the probability of getting an exact value. For example, you know that the probability of X = 4 (4 red balls). But, what if the house/host wants to know the probability of getting $\leq$ 3 red balls, as the house knows that for $\leq$ 3 red balls, the players will lose and it will make money?

Sometimes, talking in terms of 'less than' is more useful. This can be obtained by calculating the cumulative probabilities for each random variable. Let's first take a look at the observed probability table given below.

x	**F(x) = P(X = x)**
0	0.0256
1	0.1536
2	0.3456
3	0.3456
4	0.1296

The cumulative probability can be obtained by simply adding the previous probability value to the current one, i.e., F(x) = P(X) + P(X - 1). The cumulative probability distribution table should be as shown below.

x	**F(x) = P(X ≤ x)**
0	0.0256
1	0.0.256 + 0.1536 = 0.1792
2	0.1792 + 0.3456 = 0.5248
3	0.5248 + 0.3456 = 0.8704
4	0.8704 + 0.1296 = 1

Hence, the cumulative probability of X, denoted by F(x), is defined as the probability of the variable that is less than or equal to x. In mathematical terms, you would write cumulative probability as $F(x) = P(X \leq x)$. For example, $F(4) = P(X \leq 4)$ and $F(3) = P(X \leq 3)$.

However, this is not entirely correct if you consider discrete distributions, because $P(X = 60)$ would also get subtracted from the value $P(X <= 65)$ when you evaluate $P(X \leq 65) - P(X \leq 60)$. The assumption that we made, that both $P(X = 60)$ and $P(X = 65)$ would be equal to 0, is necessary for the calculation to hold true. However, an important concept that you will learn further is that the 'weight' variable is generally considered to be continuous and not discrete, and for continuous variables, the value of $P(X=x)$, where X is a random variable and x is a value, is always 0. So, for cumulative distribution tables where the probabilities of individual values are not given, you can simply use a similar analogy to calculate the probability between the two values.

SUMMARY

Continuous Probability Distributions

In this session, you learnt about continuous probability distributions. You learnt how the probability of a continuous variable is expressed and how it is different from the way the probability of a discrete variable is expressed. You also gained an understanding of normal distributions.

Continuous Random Variables

In this segment, you learnt about continuous random variables. In many situations, it is not possible for the observation to be an exact value or a discrete value. Sometimes, you might just know about the range in which your observation lies. Therefore, your answer will most likely be in the terms of ‘less than’ or ‘more than’ instead of being ‘exactly equal’ to a value. Such observations are known as continuous observations.

Continuous random variables are variables that deal with continuous values. Unlike discrete random variables, which can take a fixed number of values, a continuous random variable can take infinite values. For a continuous random variable, the probability is not defined for a single value; rather, it is defined for a range of values. The formula for probability is given below.

$$
\begin{aligned}
P(a \le x \le b) &= \int_{b}^{a} f(x)dx \\
&= \int_{-\infty}^{a} f(x)dx \; - \; \int_{-\infty}^{b} f(x)dx \\
&= F(a) - F(b)
\end{aligned}
$$

Probability Density Functions

In this segment, you learnt about probability density functions. In the previous segment, you learnt the formula to calculate the probability of continuous random variables using the cumulative density function. This method can be very difficult to do at times. Therefore, you will be using the probability density functions (PDF) to calculate the probabilities of continuous variables. Let’s recall an example to understand this concept.

Problem statement: Let’s say you work in a company with 3,000 employees. You are analysing the average time taken by the employees to commute to the office. Now, suppose you want to find the probability that it takes 35 minutes for an employee to reach the office. What is the probability going to be?

Question: Even though you do not have the probability table for this analysis, you can answer the question. The probability of a person reaching office in 35 minutes will be approximately 0. Why?

Solution: This is because time is a continuous variable and to attain exactly 35 minutes is highly unlikely. Someone may reach in 35 minutes 10 seconds, while someone else may reach in 35 minutes 43 seconds, and so on. Therefore, it is unlikely to attain exactly 35 minutes as the commute time.

This is why you need to consider different ranges of values while dealing with continuous variables. The table provided below shows the probabilities and their respective cumulative values.

x(Commute Time) (in Minutes)	Probability: F(x) = P(X = x)	Cumulative Probability: F(x) = P(X ≤ x)
20–25	0.15	0.15
25–30	0.20	0.35
30–35	0.30	0.65
35–40	0.20	0.85
40–45	0.15	1.00

A CDF, or a cumulative distribution function, is a distribution that plots the cumulative probability of X against X. If you plot a graph of the cumulative distribution, you will observe the following:

1. It is a monotonically non-decreasing function. The reason behind it is that it is a cumulative density function where you have added non-negative numbers and the sum of non-negative numbers does not decrease.
2. It should always reach the value 1.

From the CDF, you can derive a probability density function (PDF). A PDF, or a probability density function, however, is a function in which the area under the curve gives you the cumulative probability. The main difference between the cumulative probability distribution of a continuous random variable and that of a discrete one lies in the way you plot them. While a continuous variable’s cumulative distribution is a curve, a distribution for discrete variables looks more like a bar chart. The reason for this difference is that for discrete variables, the cumulative probability does not change very frequently.

You also learnt about the mean and variance of continuous variables in this segment. They can be calculated using the formulas given below:

- $E(x) = \int_{-\alpha}^{\alpha} xf(x)dx$
- $Var(x) = \sigma^2 = \int_{-\alpha}^{\alpha} (x - \mu)^2 f(x)dx$

Normal Distribution

A normal distribution is one of the most commonly occurring distributions in nature. It has the following two properties:

1. It is symmetric about its mean(μ).
2. It is dissymmetric about the median and the mode.

Normally distributed data follows the 1-2-3 rule. This rule states that there is a:

1. 68% probability of the variable lying within 1 standard deviation of the mean,
2. 95% probability of the variable lying within 2 standard deviations of the mean, and

3. 99.7% probability of the variable lying within 3 standard deviations of the mean.

This is similar to this case: If you buy a loaf of bread every day and measure it, where the mean weight = 100 grams (g) and the standard deviation = 1 g, then:

1. For 5 days every week, the weight of the loaf that you bought that day will be within 99 g (100 - 1) and 101 g (100 + 1).
2. For 20 days every 3 weeks, the weight of the loaf that you bought that day will be within 98 g (100 - 2) and 102 g (100 + 2).
3. For 364 days every year, the weight of the loaf that you bought that day will be within 97 g (100 - 3) and 103 g (100 + 3).

A lot of naturally occurring variables are normally distributed. For example, the heights of a group of adult men would be normally distributed. To try this out, we took the heights of 50 male employees at the upGrad office and then plotted the probability density function using that data.

Standard Normal Distribution

There are different types of normal distributions. This segment will focus on the standard normal distribution, which is one of the simplest and most important types. A standard normal distribution($\mu = 0$, $\sigma = 1$) should follow the following criteria:

1. Mode = 0
2. PDF should become very small near -3 and +3.

A standardised random variable is an important parameter. It is given by $Z = (X - \mu)/\sigma$. Basically, it tells you how many standard deviations away from the mean your random variable is. Alternatively, you can use the following equation to find the cumulative probability.

$$F(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{\frac{-t^2}{2}} dt$$

SUMMARY

Central Limit Theorem

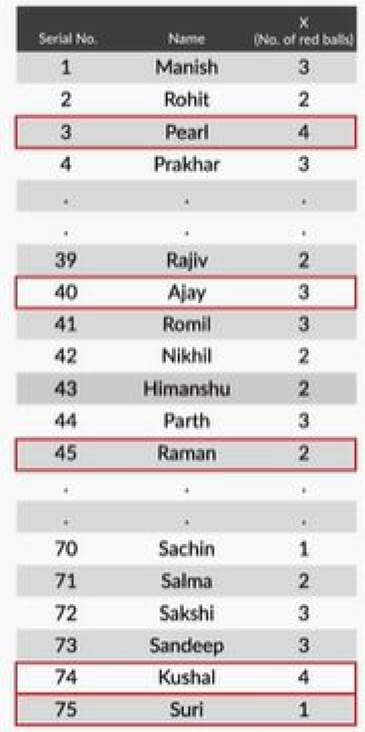
In this session, you got an understanding of what a sample is and why it is so error-prone. You also learnt how to quantify this error, which is made during sampling, using a popular theorem in statistics called the **Central Limit Theorem**.

Sampling Distribution

In this session, you learnt how to estimate results for an analysis where the size of the population is very large and it is not possible to include the entire population in the experiment.

This can be done using a **sampling distribution.** The properties of sampling distributions can help you estimate the population mean from the sample mean. Remember the red ball game, where you used the same game to understand sampling distributions.

So, you need to find the average number of red balls obtained by 75 people in the game. As you already know, the exact calculated value of the mean is 2.385; however, for this example, let's assume that the mean is not known to us. Now, the outcomes of the game for 75 people are given below. Out of 75 people, 5 people are chosen at random and their mean is calculated.



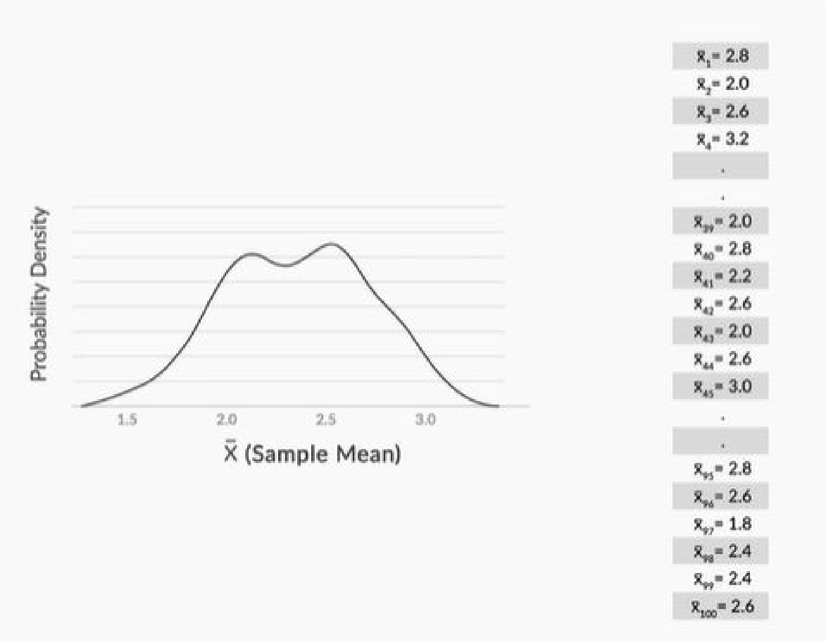
Serial No.	Name	X (No. of red balls)
1	Manish	3
2	Rohit	2
3	Pearl	4
4	Prakhar	3
.	.	.
.	.	.
39	Rajiv	2
40	Ajay	3
41	Romil	3
42	Nikhil	2
43	Himanshu	2
44	Parth	3
45	Raman	2
.	.	.
.	.	.
70	Sachin	1
71	Salma	2
72	Sakshi	3
73	Sandeep	3
74	Kushal	4
75	Suri	1

;

The people marked in red are the ones randomly selected. The mean of the red balls obtained by them is given as follows:

Mean = 4 + 3 + 2 + 4 + 15 = 2.8

When you choose different samples, the mean changes. After conducting this analysis for a large number of such samples, you will obtain several such means. The image given below shows 100 such means and their distribution.



So, the sampling distribution, specifically the sampling distribution of the sample means, is a probability density function for the sample means of a population. The sampling distribution's mean is denoted by $\bar{\mu}_x$, as it is the mean of the sampling means.

Properties of Sampling Distribution

In this segment, you learnt about some properties of sampling distributions. In the previous segment, you solved the red ball game for a sample of 5 balls and obtained the graph shown above. The mean of the sample came out to be 2.348 ($\bar{\mu}_x$), whereas the mean of the experiment when conducted with a large number of people in the initial sessions of this module was equal to 2.385 (μ). As you can see, these values lie close to each other. This brings us to the two properties of sampling distribution:

1. As the number of samples increases, the sample mean becomes approximately equal to the population mean, i.e.,

   Sampling distribution's mean ($\bar{\mu}_x$) = Population mean (μ), given that the number of samples is huge.

2. Higher the standard deviation, flatter is the curve and further apart are the sample means. i.e.,

   Sampling distribution's standard deviation (Standard error) = $\sigma/\sqrt{n}$, where σ is the population's standard deviation and n is the sample size.

In conclusion, if the standard deviation is high, the means lie far from the population mean, and the inference that you make about the population mean based on the sampling mean will not be very accurate, which will account for an error. This is why the sampling distribution's standard deviation is also known as the **standard error.**

Central Limit Theorem

In this session, you learnt about the **central limit theorem (CLT)**. According to this theorem, for any kind of data, given that a high number of samples has been taken, the following properties hold true:

1. Sampling distribution's mean ($\bar{\mu}_x$) = Population mean (μ)
2. Sampling distribution's standard deviation (standard error) = $\sigma/\sqrt{n}$
3. For $n \geq 30$, the sampling distribution becomes a normal distribution

You also demonstrated this theorem using Python and verified its properties.

Estimating Mean Using CLT

In this segment, you learnt how to compute the mean using the central limit theorem. For this purpose, the employee commute time example was reconsidered where you tried to estimate the mean commute time of 30,000 employees of an organisation by taking a small sample of 100 employees and finding their mean commute time. You know the following: Sample mean, $\bar{X}$ = 36.6 minutes; standard deviation, S = 10 minutes; and sample size, n = 100.

Recall that the population mean, i.e., the daily commute time of all 30,000 employees is given by μ = 36.6 (sample mean) ± some margin of error. This margin is calculated using the CLT.

Using the information available above, you computed the mean of the entire population.

1. Sampling distribution's mean ($\bar{\mu}_x$) = Population mean(μ)
2. Sampling distribution's standard deviation (S.E.) = $\sigma/\sqrt{n} = S/\sqrt{n} = 10/\sqrt{100} = 1$
   Since, S.E. = 1, you can approximate the sampling distribution to a normal distribution.

The next step is to find the probability that the sample mean lies between (μ+2, μ−2) i.e., the error is ±2. Using the 1-2-3 rule of normal distribution, you will get the probability of 95.4% i.e., **P(μ−1 < 36.6 < μ+2) = 95.4%**

Hence, you can conclude the following: **P(μ−1 < 36.6 < μ+2) = P(36.6−2 < μ < 36.6+2) = 95.4%.**

Therefore, you can say that there is a 95.4% probability that μ lies between 36.6±2. In other words, you can say that you are 95.4% confident that the error made in the estimation is less than or equal to 2. This probability is known as the **confidence probability,** and the maximum error is known as the **margin of error.** The range in which the mean lies is known as the **confidence interval**.

Let's suppose you have a sample size n, mean $\bar{X}$ and standard deviation S. Now, the y% confidence interval (i.e., the confidence interval corresponding to a y% confidence level) for μ would be given as follows:

Confidence interval = $$(\bar{X} - \frac{Z^*S}{\sqrt{n}}, \bar{X} + \frac{Z^*S}{\sqrt{n}})$$ **,**

where, Z* is the Z-score associated with a y% confidence level. In other words, the population mean and the sample mean differ by a margin of error given by $\frac{Z^*S}{\sqrt{n}}$.

SUMMARY

Sampling Methods

In this session, you learnt about the process of acquiring this sample data. It is important to understand this process because if your sample data is not acquired properly, even the correct application of the central limit theorem will not give you correct answers.

Types of Errors

While working on inferential statistics, you may encounter some errors because the process involves making many assumptions. You may encounter two types of errors, which are as follows:

1. Random error
2. Systematic error

While random errors occur due to the randomness of the sampling process, systematic errors occur due to the bias present in the data sample. Therefore, in order to reduce the chances of making these errors, you need to select the correct sampling method.

Types of Sampling

There are four different types of sampling methods. Let's revise them one by one.

1. **Random sampling**: In this method, people in the sample are selected at random. *For example,* you need to determine the average internet usage per person in a country. You can simply put chits with the names of all citizens of that country in a bowl and pick 100 names at random, and then calculate the average internet usage of these 100 citizens.

2. **Stratified sampling:** In this method, people are first divided into subgroups and then randomly selected from those subgroups. However, this is done in such a way that the final sample has the same proportions of the subgroups as the entire population. *For example,* given that 70% of Indians live in rural areas and 30% of Indians live in urban areas, you need to determine the average internet usage per person in India. In this case, you would need to put chits with the names of Indians from rural areas in Hat A and those with the names of Indians from urban areas in Hat B. You would then need to pick 70 names out of Hat A and 30 names out of Hat B.

3. **Volunteer sampling**: In this method, your sample is composed of people who want to volunteer for the survey. *For example,* you need to determine the average internet usage per person in India. You could ask people to take an online survey containing questions about how often/much they use the internet. You could ask them the same question through a telephonic survey.

4. **Opportunity sampling**: In this method, the people around and close to the surveyor form the sample space. *For example,* you need to determine the average internet usage per person in India. You could just ask the 100 people who are in close proximity to you about their internet usage.

In order to decide which sampling method to use in any given situation, you need to identify your target audience and understand the kind of results you are looking for. You can do this by considering the following factors:

1. If accuracy is your main concern, you should try to avoid using volunteer and opportunity sampling methods. Even though these methods are quite useful in performing EDA, they are highly biased and may lead to inaccurate results.
2. The stratified sampling method is best suited for obtaining accurate results by reducing errors.
3. Alternatively, you can also reduce bias after using a particular sampling method with the help of statistical techniques such as reweighting.

Applications of Sampling Methods

The four typical scenarios in which sampling is generally used are as follows:

1. **Market research**: Suppose your company wants to launch a product whose usage depends on people having a decent internet connection, as is in the case of Hotstar and Netflix. Before launching such a product, you need to understand the potential market size for it. For this, you need to conduct a survey and, based on the data of the results, infer parameters such as average data usage and willingness to adopt new technologies for the entire population.

2. **Marketing campaign efficacy**: Suppose you work for a company (such as Hotstar or Netflix) and want more people to move from your competitors' platforms to your platform. You decide to launch a marketing campaign to get people to make the shift to your platform. How would you structure this marketing campaign? What should be the budget of the campaign? Which strategy should you use (free membership for a week, lower membership fees for a few weeks, etc.)? You can use the data from your past marketing campaign and your knowledge of sampling techniques to make these decisions.

3. **Pilot testing**: Let's consider the example mentioned above for point 2. Suppose you have conducted the required market research and developed your product. Now, before launching your product in the market, you want to conduct a trial run. For this, you can use a method called **pilot test**. In a pilot test, instead of launching your product in the market in all its glory, you launch it partially to a few people. These people test your product and help you decide whether it is good enough for a complete launch.

4. **Quality control**: This is a manufacturing-centred application. Suppose your company produces 10 million smartphones annually, which means that about 30,000 smartphones are manufactured every day. In such a situation, quality assurance (QA) becomes the most important function. Since it would be difficult to check all 30,000 phones every day, your company 'samples' a few phones and makes decisions based on the quality of these samples.

upGrad

Disclaimer: All content and material on the upGrad website is copyrighted material, either belonging to upGrad or its bonafide contributors and is purely for the dissemination of education. You are permitted to access, print and download extracts from this site purely for your own education only and on the following basis:

- You can download this document from the website for self-use only.
- Any copy of this document, in part or full, saved to disk or to any other storage medium, may only be used for subsequent, self-viewing purposes or to print an individual extract or copy for non-commercial personal use only.
- Any further dissemination, distribution, reproduction, copying of the content of the document herein or the uploading thereof on other websites or use of the content for any other commercial/unauthorised purposes in any way which could infringe the intellectual property rights of upGrad or its contributors, is strictly prohibited.
- No graphics, images or photographs from any accompanying text in this document will be used separately for unauthorised purposes.
- No material in this document will be modified, adapted or altered in any way.
- No part of this document or upGrad content may be reproduced or stored in any other website or included in any public or private electronic retrieval system or service without upGrad’s prior written permission.
- Any right not expressly granted in these terms are reserved.