

UNIT-3

Random Variables and Probability Distributions:

Random variables (discrete and continuous), Probability Density Function (PDF), Probability Mass Function (PMF), and Cumulative Density Function (CDF). Discrete distributions- Uniform, Binomial, Bernoulli and Poisson distributions. Continuous Distributions- Normal distribution, Standard Normal distribution, Student's T distribution, Chi-squared distribution.

Sampling Strategies: Introduction, Simple Random sampling, Systematic sampling, Stratified sampling, Cluster sampling.

Random Variables and Probability Distributions:

Random variables are variables whose values are determined by the outcome of a random event. In other words, they are variables whose values are not fixed, but rather, determined by chance. Probability distributions, on the other hand, are mathematical functions that describe the likelihood of each possible outcome of a random variable. They are used to model the behavior of random variables and can provide important insights into various statistical phenomena.

Example 1: Coin Tossing Experiment

Suppose we conduct an experiment where we flip a coin once. We can define a random variable X as the number of heads that appear when we flip the coin. The possible values of X are 0 (no heads) or 1 (one head).

The probability distribution of X is given by:

X	$P(X)$
0	0.5
1	0.5

This is a simple example of a probability distribution called the Bernoulli distribution. It models the behavior of a binary event (heads or tails) and assigns a probability to each possible outcome.

Example 2: Rolling a Dice

Suppose we roll a six-sided dice once. We can define a random variable Y as the number that appears on the dice. The possible values of Y are 1, 2, 3, 4, 5, or 6.

The probability distribution of Y is given by:

Y	P(Y)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

This is an example of a probability distribution called the uniform distribution. It models the behavior of a random event with equally likely outcomes.

Example 3: Heights of Students

Suppose we measure the heights of 10 students in a class. We can define a random variable Z as the height of a randomly selected student. The possible values of Z are continuous and can take any value between the minimum and maximum height of the students.

A continuous function gives the probability distribution of Z called the probability density function (PDF). For simplicity, let's assume that the heights are normally distributed with a mean of 170 cm and a standard deviation of 5 cm. Then, the PDF of Z is given by:

$$f(z) = (1 / (5\sqrt{2\pi})) * e^{-((z-170)^2)/(2*5^2)}$$

This PDF describes the likelihood of each possible height. It can be used to calculate probabilities of various events, such as the probability that a randomly selected student is taller than 175 cm.

Random variables (discrete and continuous):

Random variables are variables that take on random values based on the outcome of a random event. They can be classified into two main types: discrete random variables and continuous random variables.

Discrete Random Variables: A discrete random variable takes on a countable number of distinct values. It is used to represent a situation where the outcome of an experiment is one of a finite or countably infinite number of possible outcomes.

Examples of discrete random variables include the number of heads in a series of coin flips, the number of cars passing through a traffic signal, the number of defects in a manufacturing process, and so on.

For example, let's consider the number of times a coin comes up heads when it is flipped three times. This can be modeled as a discrete random variable with the possible values of 0, 1, 2, or 3. The probability distribution of this random variable can be represented by a table or a graph that shows the probability of each possible outcome. For instance, the probability of getting 0 heads is $1/8$, the probability of getting 1 head is $3/8$, the probability of getting 2 heads is $3/8$, and the probability of getting 3 heads is $1/8$.

Continuous Random Variables: A continuous random variable takes on any value within a specified range. It is used to represent a situation where the outcome of an experiment can take on an infinite number of possible outcomes. Examples of continuous random variables include height, weight, time, temperature, and so on.

For example, let's consider the amount of rainfall in a particular area over a period of time. This can be modeled as a continuous random variable, since the amount of rainfall can take on any value within a specified range. A continuous function can represent the probability distribution of this random variable, called the probability density function (PDF), which describes the probability of the random variable taking on any value within the specified range. For instance, the amount of rainfall may be a normal distribution, which is bell-shaped and centered around the average amount of rainfall.

Discrete random variables represent situations where the outcome of an experiment can take on a finite or countably infinite number of possible outcomes. In contrast, continuous random variables represent situations where the outcome can take on any value within a specified range.

Probability Density Function (PDF):

In probability theory and statistics, a probability density function (PDF) is a function that describes the relative likelihood for a continuous random variable to take on a given value. The PDF of a continuous random variable is a non-negative function $f(x)$ where the area under the curve between any two values a and b represents the probability of the random variable taking on a value between a and b .

The PDF is defined such that the total area under the curve equals 1. This means that the integral of the PDF over its entire domain is equal to 1. The PDF can be used to calculate probabilities of events, such as the probability that a continuous random variable takes on a value between two given points.

One common example of a PDF is the normal distribution, also known as the Gaussian distribution. The PDF of a normal distribution is a bell-shaped curve and is given by the formula:

$$f(x) = (1/\sigma\sqrt{2\pi}) e^{-(x-\mu)^2/(2\sigma^2)}$$

where μ is the mean of the distribution, σ is the standard deviation, and e is the mathematical constant approximately equal to 2.718.

Another example of a PDF is the exponential distribution, often used to model the time between events in a Poisson process. The formula gives the PDF of an exponential distribution:

$$f(x) = \lambda e^{-\lambda x}$$

where λ is the rate parameter, which specifies the average number of events per unit time.

A probability density function is a function that describes the relative likelihood of a continuous random variable taking on a given value. The PDF is a normalised non-negative function such that the total area under the curve is equal to 1. It can be used to calculate probabilities of events and is an important concept in probability theory and statistics.

Probability Mass Function (PMF)

Probability Mass Function (PMF): In probability theory, a probability mass function (PMF) is a function that gives the probability of a discrete random variable taking on a specific value. It is defined as the probability of each possible outcome of a random variable, represented as a histogram or a bar chart.

For example, consider a fair six-sided die. The PMF for this die is given by:

x	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6

Here, X is the random variable that represents the outcome of rolling the die, and the PMF gives the probability of each possible outcome.

Cumulative Density Function (CDF): The cumulative distribution function (CDF) is the probability distribution function that gives the probability that a random variable X is less than or equal to a certain value x . It is defined as the sum of the probabilities of all outcomes less than or equal to x .

For example, consider the PMF of the fair six-sided die above. The CDF for this die is:

x	1	2	3	4	5	6
F(x)	1/6	1/3	1/2	2/3	5/6	1

Here, $F(x)$ represents the cumulative probability that the random variable X is less than or equal to x . For example, the probability that X is less than or equal to 3 is $1/2$, because $F(3) = 1/2$.

The probability mass function (PMF) gives the probability of each possible outcome of a discrete random variable. In contrast, the cumulative distribution function (CDF) gives the probability that a random variable is less than or equal to a certain value. Both are important concepts in probability theory and statistics and are used to calculate probabilities of events and make predictions based on data.

Discrete distributions- Uniform, Binomial, Bernoulli and Poisson distributions:

Discrete probability distributions are used to model random variables that can only take on a countable number of values. This will explain four common discrete probability distributions: Uniform, Binomial, Bernoulli, and Poisson.

1. Uniform distribution: A uniform distribution is a probability distribution where each value in a finite or infinite range has an equal probability of occurring. For example, when a fair six-sided die is rolled, the probability of getting each number from 1 to 6 is the same, or $1/6$.
2. Binomial distribution: The binomial distribution models the number of successes in a fixed number of independent trials, where each trial has a binary outcome (success or failure). The probability of success on each trial is denoted as p , and the probability of failure is denoted as $q = 1 - p$. For example, the number of heads obtained when flipping a fair coin 10 times is a binomial random variable. The probability mass function (PMF) of the binomial distribution is given by:

$$\text{PMF}(x) = \binom{n}{x} * p^x * q^{(n-x)}$$

where n is the total number of trials, x is the number of successful trials, and $\binom{n}{x}$ is the binomial coefficient.

3. Bernoulli distribution: The Bernoulli distribution is a special case of the binomial distribution, where there is only one trial. It models the probability of success or failure in a single trial. For example, the outcome of a single coin flip can be modeled using a Bernoulli random variable. The PMF of the Bernoulli distribution is given by:

$$\text{PMF}(x) = p^x * (1-p)^{(1-x)}$$

where x can only take on the values 0 or 1.

4. Poisson distribution: The Poisson distribution models the number of events occurring in a fixed time interval or space, given a certain rate of occurrence. For example, the number of car accidents occurring in a given hour on a particular road can be modeled using a Poisson random variable. The PMF of the Poisson distribution is given by:

$$\text{PMF}(x) = \frac{e^{(-\lambda)} * \lambda^x}{x!}$$

where λ is the average rate of occurrence, and x is the number of occurrences in the fixed time interval or space.

Continuous Distributions- Normal distribution, Standard Normal distribution, Student's T distribution, Chi-squared distribution:

Continuous probability distributions are used to model random variables that can take on any value within a continuous range. This will explain four common continuous probability distributions: Normal, Standard Normal, Student's T, and Chi-squared.

1. Normal distribution: The normal distribution is a bell-shaped curve that is symmetrical around its mean, and its standard deviation determines its spread. Many natural phenomena, such as heights and weights of individuals, follow a normal distribution. The probability density function (PDF) of the normal distribution is given by:

$$\text{PDF}(x) = (1 / (\sigma * \sqrt{2\pi})) * e^{-(x - \mu)^2 / (2 * \sigma^2)}$$

where μ is the mean of the distribution, σ is the standard deviation, and π is the mathematical constant.

2. Standard Normal distribution: The standard normal distribution is a special case of the normal distribution, where the mean is zero and the standard deviation is one. It is often used as a standard for comparison and can be denoted by the letter Z . The PDF of the standard normal distribution is given by:

$$\text{PDF}(x) = (1 / \sqrt{2\pi}) * e^{-x^2 / 2}$$

3. Student's T distribution: The Student's T distribution is used when the sample size is small or the population standard deviation is unknown. It is used in hypothesis testing, confidence intervals, and in the construction of prediction intervals. The PDF of the Student's T distribution is given by:

$$\text{PDF}(x) = (\Gamma((v+1)/2) / (\sqrt{\pi * v} * \Gamma(v/2))) * (1 + x^2/v)^{-(v+1)/2}$$

where v is the degrees of freedom parameter and Γ is the gamma function.

4. Chi-squared distribution: The Chi-squared distribution is used in hypothesis testing, goodness-of-fit tests, and in the construction of confidence intervals

for the variance of a normally distributed population. The PDF of the Chi-squared distribution is given by:

$$\text{PDF}(x) = (1 / (2^{(k/2)} * \Gamma(k/2))) * x^{(k/2 - 1)} * e^{(-x/2)}$$

where k is the degrees of freedom parameter and Γ is the gamma function.

Continuous probability distributions are used to model random variables that can take on any value within a continuous range. Here are a few examples of how continuous distributions are used in various fields:

1. Normal distribution: The normal distribution is used in many fields, including finance, engineering, and the natural sciences. For example, in finance, stock prices can be modeled using a normal distribution, while in engineering, the heights of individuals can be modeled using a normal distribution.
2. Standard Normal distribution: The standard normal distribution is used as a standard for comparison and can be used to calculate probabilities for other normal distributions. For example, suppose a sample of heights is normally distributed with a mean of 65 inches and a standard deviation of 3 inches. In that case, we can use the standard normal distribution to calculate the probability that an individual's height is between 62 and 68 inches.
3. Student's T distribution: The Student's T distribution is commonly used in hypothesis testing and confidence intervals when the sample size is small or the population standard deviation is unknown. For example, in medical research, a small sample of patients may be used to test the effectiveness of a new treatment, and the Student's T distribution can be used to test whether the treatment has a statistically significant effect.
4. Chi-squared distribution: The Chi-squared distribution is commonly used in hypothesis testing, goodness-of-fit tests, and in the construction of confidence intervals for the variance of a normally distributed population. For example, in quality control, a manufacturer may use a Chi-squared distribution to test whether the variance of a product's weights is within acceptable limits.

Sampling Strategies: Introduction, Simple Random sampling, Systematic sampling, Stratified sampling, Cluster sampling:

Sampling is the process of selecting a subset of individuals or units from a larger population in order to estimate characteristics or parameters of the population. Several sampling strategies can be used depending on the research question, available resources, and the nature of the population. In this response, we will discuss the following sampling strategies: simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

1. **Simple Random Sampling:** Simple random sampling is a method where each individual or unit in the population has an equal chance of being selected for the sample. This is achieved by selecting individuals or units at random from the population. For example, if a researcher wants to conduct a study on the average age of individuals in a town, they could randomly select a certain number of individuals from a list of all residents in the town. Simple random sampling can be done with or without replacement.
2. **Systematic Sampling:** Systematic sampling is a method where individuals or units are selected at regular intervals from the population. This is done by selecting a starting point at random and then selecting every k th individual or unit from the population. For example, if a researcher wants to conduct a study on the weight of bags of rice in a warehouse, they could select every 10th bag of rice from a line of bags on a conveyor belt.
3. **Stratified Sampling:** Stratified sampling is a sampling method where the population is divided into groups or strata based on specific characteristics and a sample is selected from each stratum. This is done to ensure that the sample is representative of the population for the characteristics being studied. For example, if a researcher wants to conduct a study on the academic performance of students in a school, they could divide the student population into strata based on grade level and then randomly select a certain number of students from each grade level stratum.
4. **Cluster Sampling:** Cluster sampling is a sampling method where the population is divided into clusters, and a random sample of clusters is selected for the study. Within each selected cluster, all individuals or units are included in the sample. This is often used when obtaining a complete list of individuals or units in the population is difficult or impractical. For example, if a researcher wants to conduct a study on the prevalence of a disease in a country, they could divide the country into regions and randomly select a certain number of regions for the study. Each selected region would include all individuals with the disease in the sample.