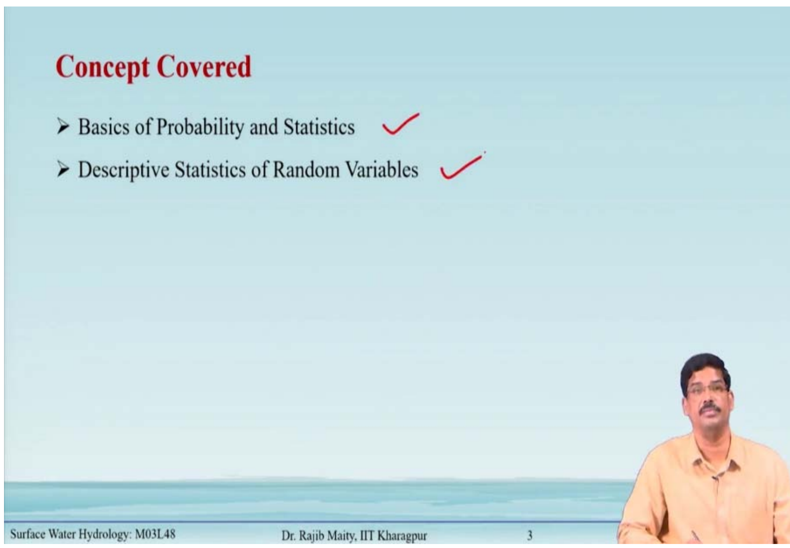
Surface Water Hydrology
Professor Rajib Maity
Department of Civil Engineering
Indian Institute of Technology, Kharagpur
Lecture 48
Basic Concepts of Probability and Statistics

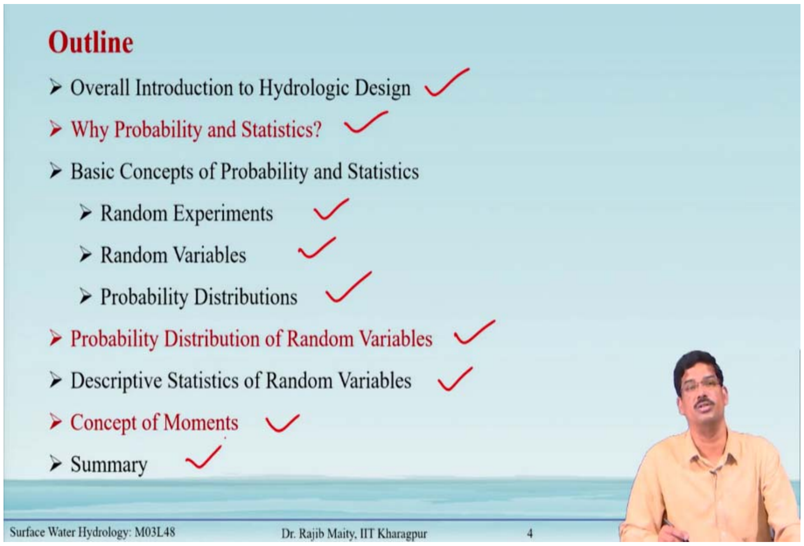
In this week 10, we will be mainly focusing on hydrologic frequency analysis. And to this particular lecture, we will learn some basic concepts of probability and statistics which are very much essential for hydrologic frequency analysis as well as for hydrologic design in general.

(Refer Slide Time: 00:59)



Today's lecture will try to learn the concept of the basics of probability and statistics and some descriptive statistics of random variables.

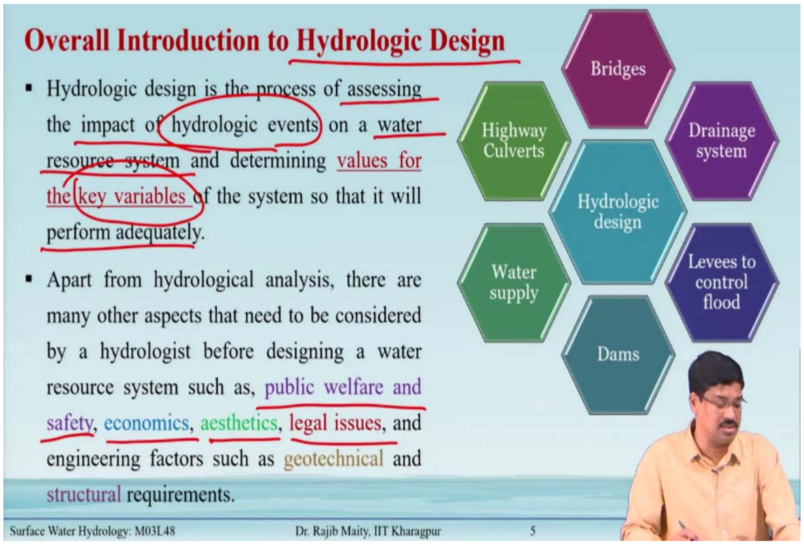
(Refer Slide Time: 01:26)



The outline of this lecture goes like this, first of all, we will give one overall introduction to hydrologic design, why it is, what the different aspects of it are, and why is it important how we can utilize our known knowledge base to these different hydrologic design perspectives.

Under these basics of probability and statistics, we will be learning the concept of a random experiment, random variable, and different probability distributions. And then probability distributions of random variables and also the descriptive statistics that we can extract from the sample or for a particular random variable. The concept of moments is also another important thing that we need to know. Finally, we will summarize what we have learned for this one.

(Refer Slide Time: 02:45)



Overall Introduction to Hydrologic Design

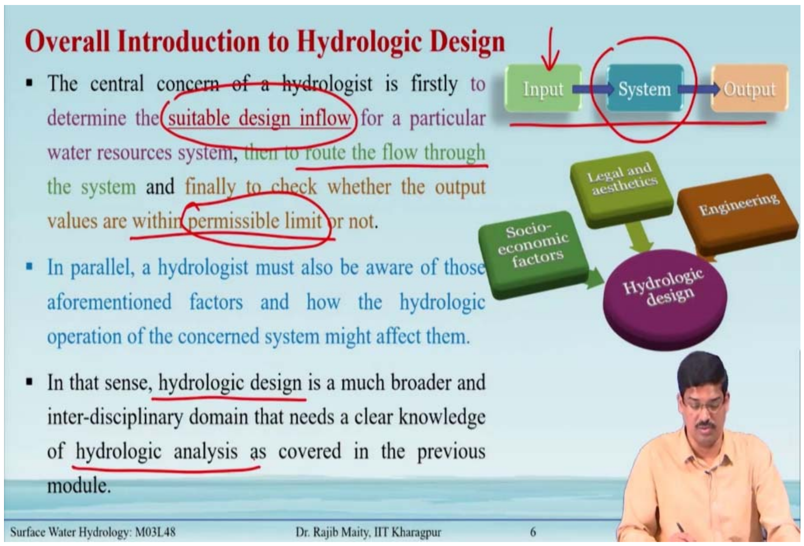
It is a process of assessing the impact of hydrologic events on a water resource system and determining the values for the key variables of the system so that it will perform adequately.

For any design purpose, we need to identify some of the key variables, for example, looking for a culvert, then have to know what is the maximum discharge that it can it should pass through. And then go for drainage, then we should know what should be the peak discharge that it should take. If we look for some water storage structure like a dam and also how much storage volume is required.

Those are the key variables that we take a decision and the impact of the hydrologic event. So, when we talk about the impact of the hydrologic event, that means we have to identify not only from the historical data but also what is this possible design value that we should start with. Otherwise, that particular value to first of all to identify that particular value itself is a task that we have to go through initially.

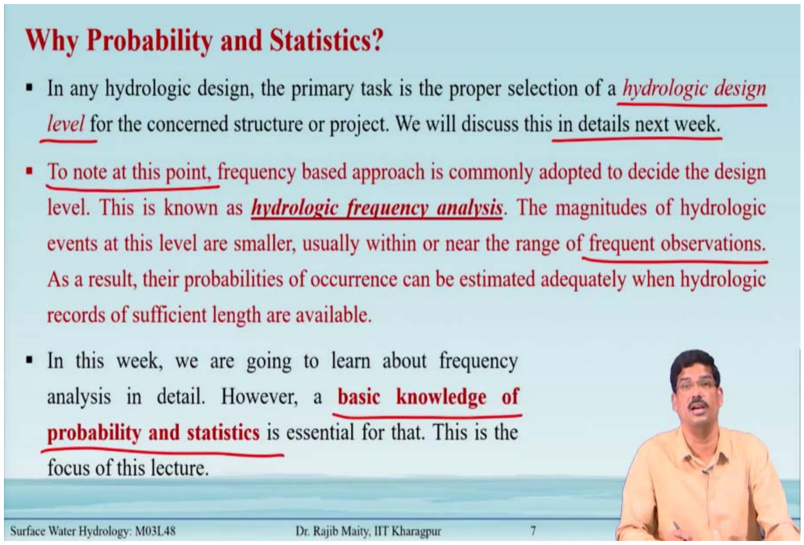
Some examples of hydrologic design are bridges, water supply, dams, and levees to control floods and protect some areas of importance. Not only that, but there are other issues are also there which come from the other particular perspectives, for example, the public welfare and safety is also one of the important things then come economics, how economic or economically viable that particular project is. Aesthetics is also another important factor. Some legal issues also can be associated with it such as geotechnical or the structural requirement and all.

(Refer Slide Time: 05:01)



The central concern of a hydrologist is first to determine the suitable design info or suitable design value for any particular key variable for a particular water resources system, then route the entire flow through the system and finally check whether the outflow values are within the permissible limit or not. Any system has to decide what input should be put in and then will check whether the system can take that input and safely produce the output without going or crossing the permissible limit at any point. And along with that, the hydrologist must also be aware of those aforementioned factors like socioeconomic factors, legal and aesthetic issues, and engineering perspectives, those things should also be there and how the hydrologic operation of the constant system might affect them that needs to be studied. This hydrologic design when we talk about it is a much broader and interdisciplinary domain that needs a clear knowledge of the hydrologic analysis course, that is required, which we have already covered in the previous module.

(Refer Slide Time: 06:31)

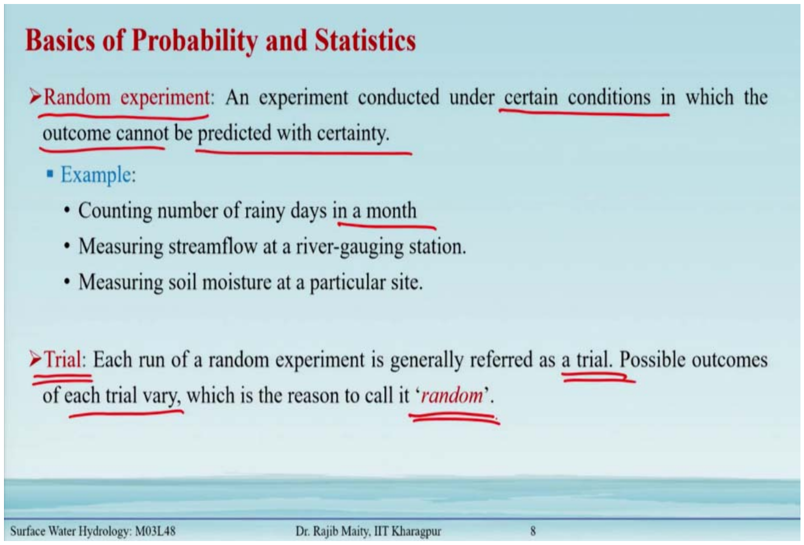


Why Probability and Statistics?

In any hydrologic design, the primary task is the proper selection of a hydrologic design level for the concerned structure or project. The frequency-based approach is commonly adopted to decide the design level and this is sometimes called the hydrologic frequency analysis. The magnitude of the hydrologic event at this level is smaller and usually within or near the range of the frequent observations that we can see in the data record itself.

As a result, their probabilities of occurrence can be estimated adequately when hydrologic records of sufficient length are available. Now, this is one of the aspects that is hydrologic frequency analysis in there are many other concepts are there that are involved in the hydrologic design, where this concepts along with this hydrologic frequency analysis, those things requires a basic knowledge of probability and statistics that is essential.

(Refer Slide Time: 08:07)



Basics of Probability and Statistics

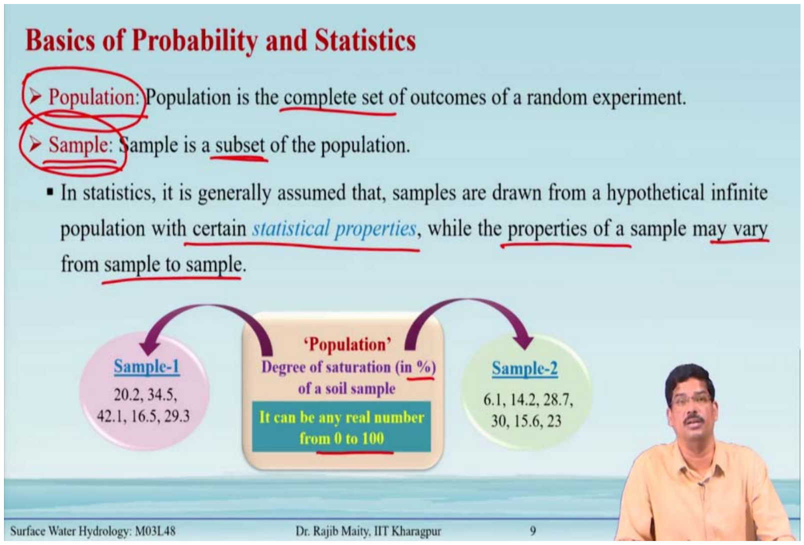
➢ Random experiment: An experiment conducted under certain conditions in which the outcome cannot be predicted with certainty.

Example:

- Counting the number of rainy days in a month
- Measuring streamflow at a river-gauging station.
- Measuring soil moisture at a particular site.

➢ Trial: Each run of a random experiment is generally referred to as a trial. Possible outcomes of each trial vary, which is the reason to call it '*random*'.

(Refer Slide Time: 09:46)

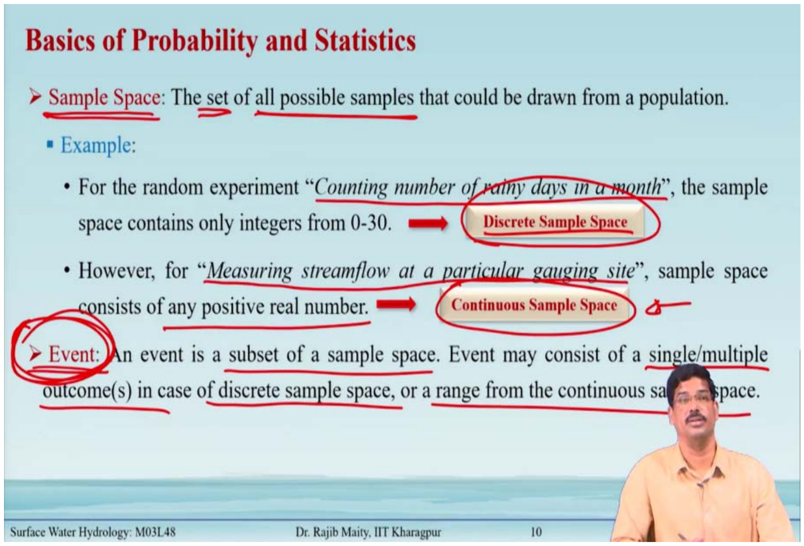


- ➢ Population: Population is the complete set of outcomes of a random experiment.
- ➢ Sample: The sample is a subset of the population.

In statistics, it is generally assumed that samples are drawn from a hypothetical infinite population with certain statistical properties, while the properties of a sample may vary from sample to sample.

Now, one example of this population is the degree of saturation of a soil sample can be any real number from 0 to 100. Now, sample means if collect it from some records example, sample1 is 20.2, 34.5, 42.1, 16.5, and 29.3 and collect another sample which will definitely have some different numbers and all.

(Refer Slide Time: 13:24)



- Sample Space: The set of all possible samples that could be drawn from a population.

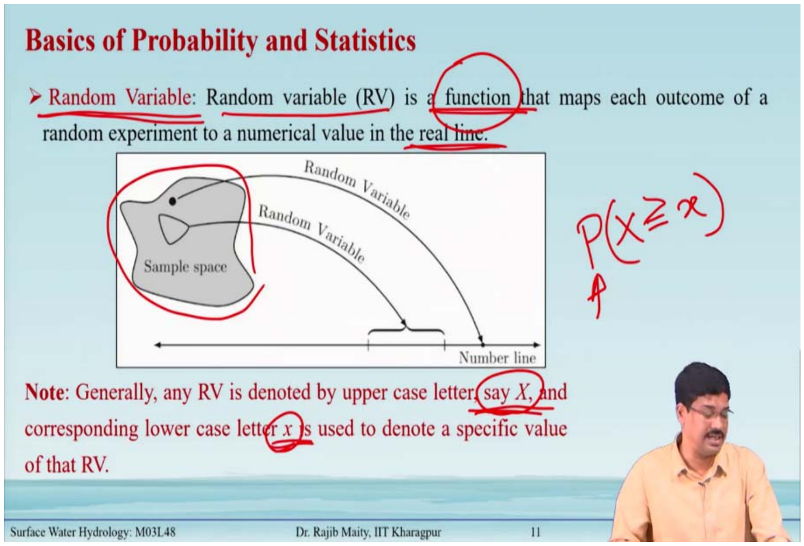
  Example:

For the random experiment "Counting number of rainy days in a month", the sample space contains only integers from 0-30. This kind of sample space is called the discrete sample space it cannot be any fraction.

However, for "Measuring streamflow at a particular gauging site", sample space consists of any positive real number. This kind of sample space is called the continuous sample space.

- Event: An event is a subset of a sample space. The event may consist of a single/multiple outcome(s) in the case of discrete sample space or a range from the continuous sample space.

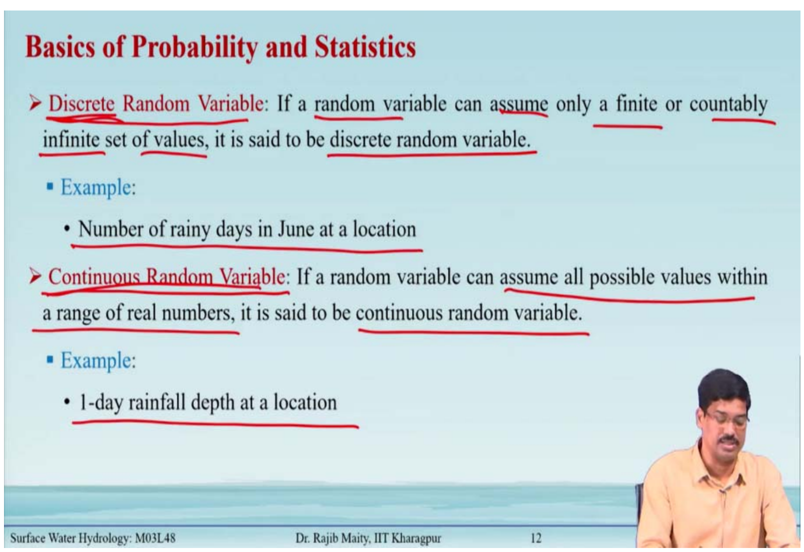
(Refer Slide Time: 16:21)



- Random Variable: Random variable (RV) is a function that maps each outcome of a random experiment to a numerical value in the real line.

It may be noted that Generally, any RV is denoted by an upper case letter, say *X*, and the corresponding lower case letter *x* is used to denote a specific value of that RV.

(Refer Slide Time: 18:20)



➢ Discrete Random Variable: If a random variable can assume only a finite or countably infinite set of values, it is said to be a discrete random variable.

Example:

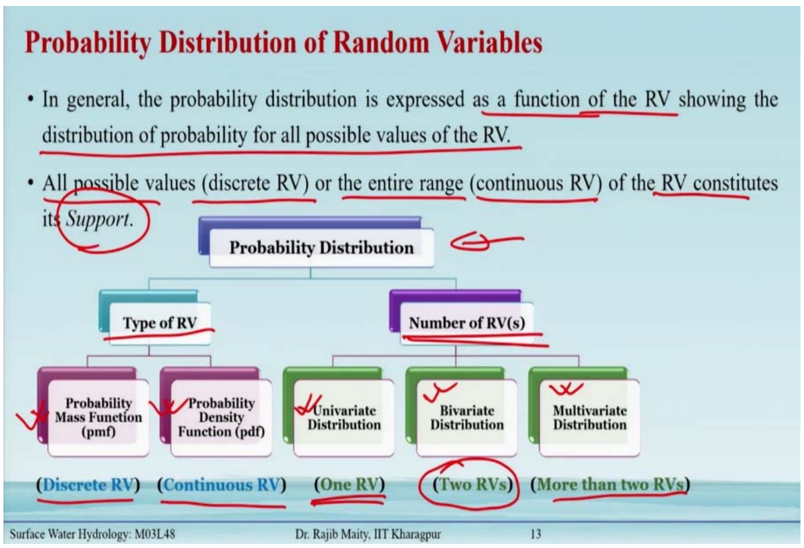
- Number of rainy days in June at a location

➢ Continuous Random Variable: If a random variable can assume all possible values within a range of real numbers, it is said to be a continuous random variable.

Example:

- 1-day rainfall depth at a location

(Refer Slide Time: 19:21)



Probability Distribution of Random Variables

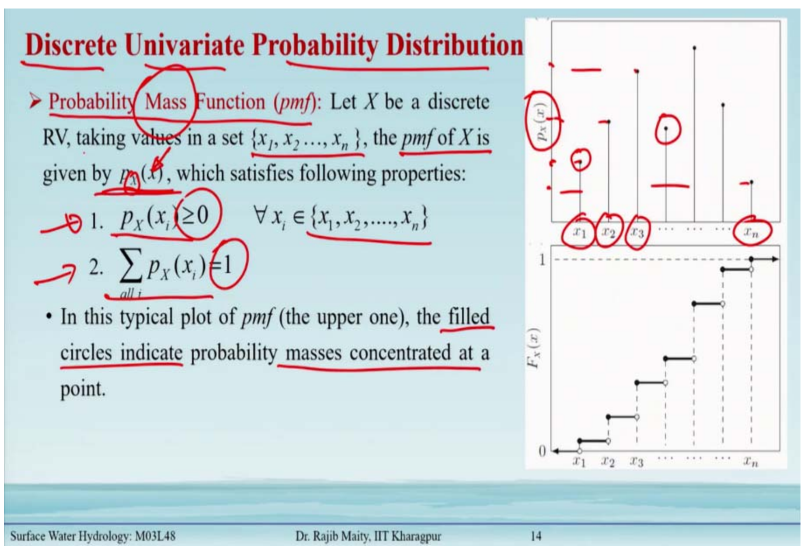
The probability distribution is expressed as a function of the random variable showing the distribution of probability for all possible values of the random variable. The probability distribution is a function of this random variable that shows the distribution of the probability.

Now, all possible values of a discrete random variable or the entire range of the random variable that is there is a term called support.

Now, so far as this for now, what we should know is that the deviation of the categorization of this probability distribution, so, first of all depending on the type of this random variable there are say two types one is the probability mass function and probability density function, so probability mass function is for the discrete random variable, the probability density function is for the continuous random variable.

So, far as the number of random variables is involved, we can say that it is univariate, bivariate, or multivariate. There is no very clear division between bivariate and multivariate, if it is more than one itself, we can say that is multivariate, but when there are two random variables, we generally give it a special name as the bivariate. And more than two random variables sometimes we can call it trivariate also or in general multivariate, but if it is the only single random variable, then we call it univariate distribution, this is the different categories of this probability distribution.

(Refer Slide Time: 21:32)



➤ Probability Mass Function (*pmf*): Let X be a discrete RV, taking values in a set $\{x_1, x_2 \ldots, x_n\}$, the *pmf* of X is given by $p_X(x)$, which satisfies the following properties:

1. $p_X(x_i) \geq 0 \qquad \forall x_i \in \{x_1, x_2, \ldots, x_n\}$

2. $\sum_{all\ i} p_X(x_i) = 1$

In fig.1 the typical plot of *pmf* (the upper one), the filled circles indicate probability masses concentrated at a point.

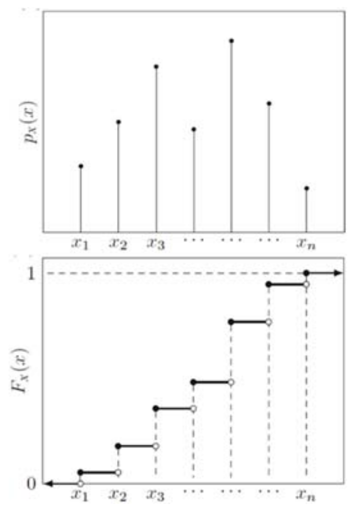
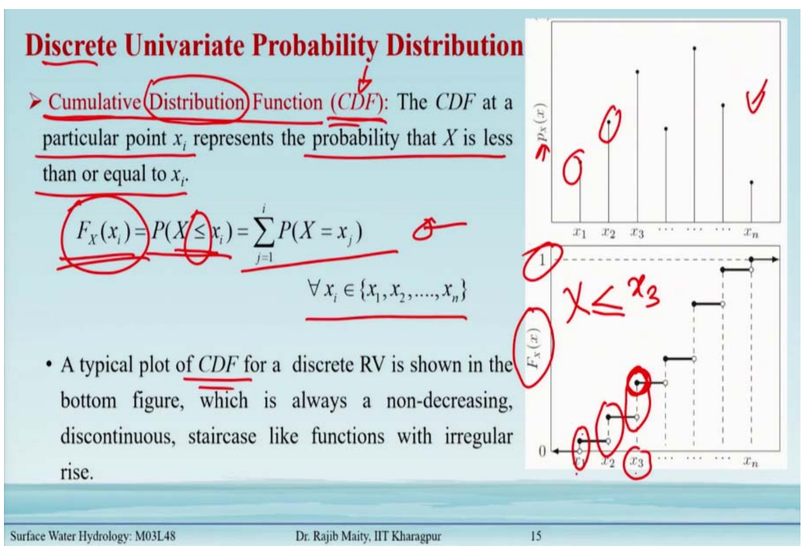


Figure 1 shows a typical plot of *pmf*

(Refer Slide Time: 23:31)



➢ Cumulative Distribution Function (*CDF*): The *CDF* at a particular point x_i represents the probability that X is less than or equal to x_i.

$$F_X(x_i) = P(X \le x_i) = \sum_{j=1}^{i} P(X = x_j)$$

$$\forall\, x_i \in \{x_1, x_2, ..., x_n\}$$

A typical plot of CDF for a discrete RV is shown at the bottom of figure2, which is always a non-decreasing, discontinuous, staircase-like function with an irregular rise.

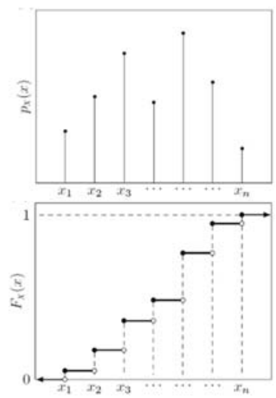
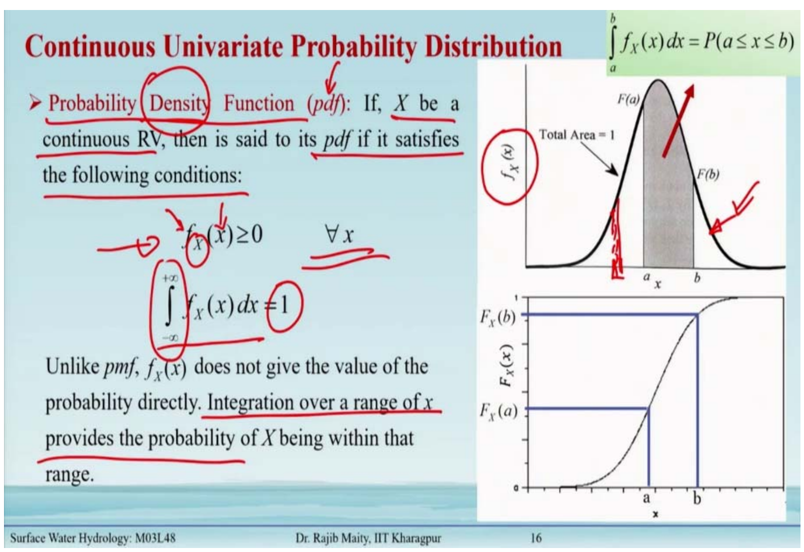


Figure 2 shows a typical plot of CDF

(Refer Slide Time: 26:33)



Continuous Univariate Probability Distribution

- Probability Density Function (*pdf*): If, X is a continuous RV, then is said to its *pdf* if it satisfies the following conditions:

$$f_X(x) \geq 0 \qquad \forall x$$

$$\int_{-\infty}^{+\infty} f_X(x)\, dx = 1$$

Unlike *pmf*, $f_X(x)$ does not give the value of the probability directly. Integration over a range of *x* provides the probability of *X* being within that range.

For a particular location it does not give that the measure of the probabilities gives only the value, but if take a very small area and integrate it over whatever the shaded area that will get, that will give you the measure of this probability, and as you can see that integration over the entire area that means, that below this curve shown in fig.3, the total area below this curve up to the positive X-axis equals to 1.

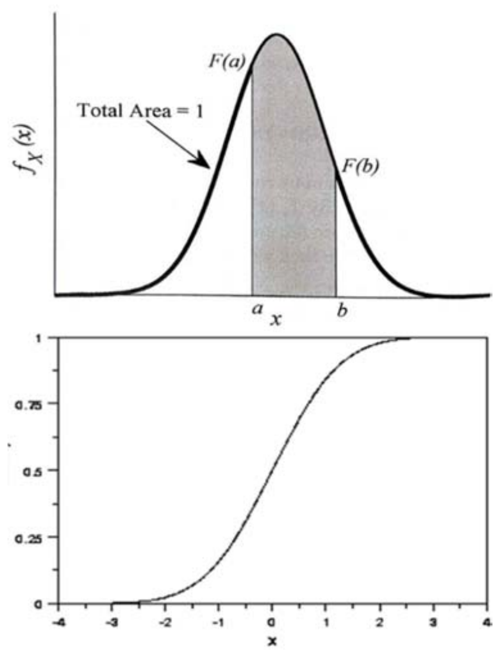
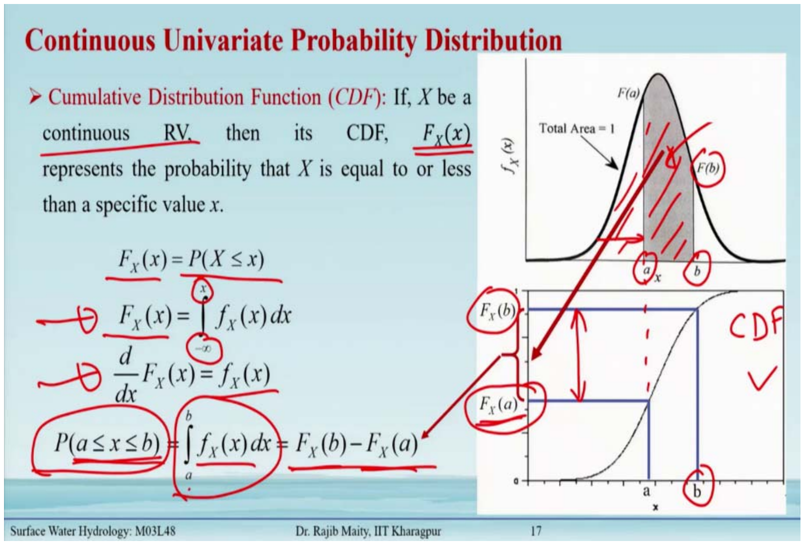


Figure 3 shows a typical pdf curve

(Refer Slide Time: 28:16)



- Cumulative Distribution Function (*CDF*): If *X* is a continuous RV, then its CDF, $F_X(x)$ represents the probability that *X* is equal to or less than a specific value *x*.

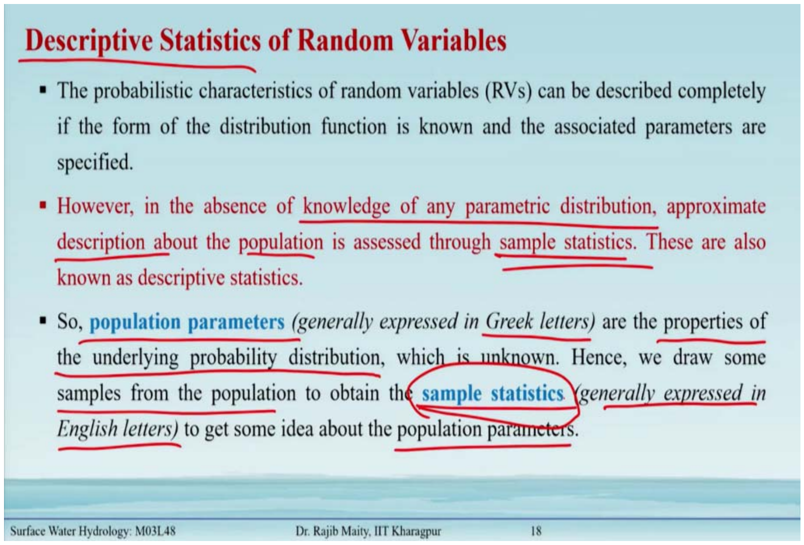
$$F_X(x) = P(X \le x)$$

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx$$

$$\frac{d}{dx}F_X(x) = f_X(x)$$

$$P(a \le x \le b) = \int_a^b f_X(x)\,dx = F_X(b) - F_X(a)$$

(Refer Slide Time: 30:29)



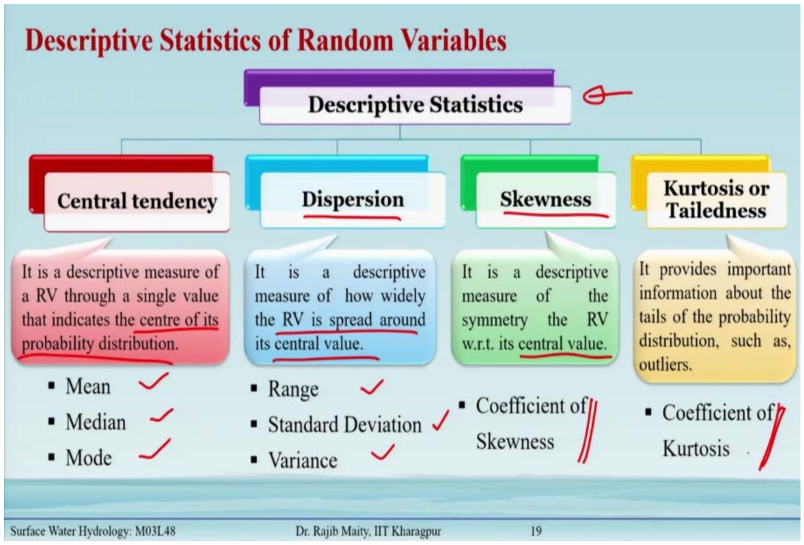
Descriptive Statistics of Random Variables

The probabilistic characteristics of random variables (RVs) can be described completely in the form of the distribution function is known and the associated parameters are specified.

However, in the absence of knowledge of any parametric distribution, an approximate description of the population is assessed through sample statistics. These are also known as descriptive statistics.

So, population parameters (generally expressed in Greek letters) are the properties of the underlying probability distribution, which is unknown. Hence, we draw some samples from the population to obtain the sample statistics (generally expressed in English letters) to get some idea about the population parameters.

(Refer Slide Time: 31:32)



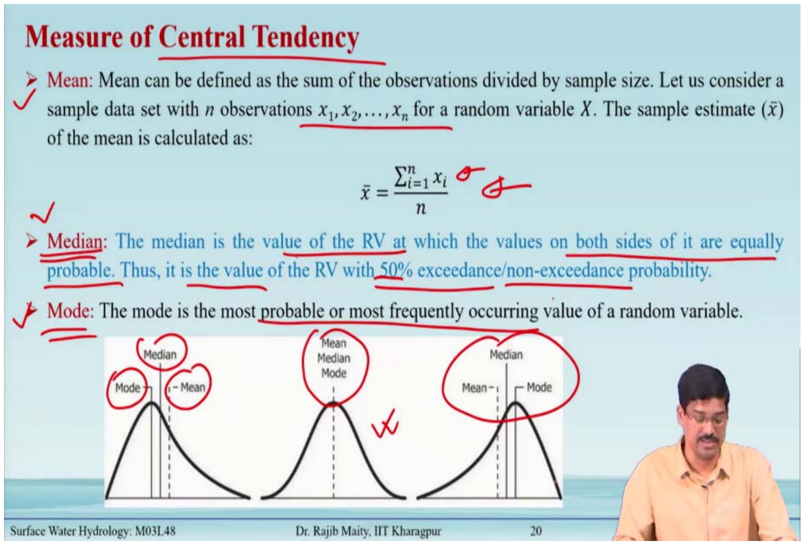
In descriptive statistics, there are a couple of categories to this one first one is the central tendency, it is a descriptive measure of the random variable through a single value that indicates the center of its probability distribution. So, remember that towards the center of his probability distribution, there are three terms of their Mean, Median, and Mode.

Similarly, the measure of dispersion when we say it is a descriptive measure of how widely or what is the range over which the random variable is spread around the central value, again there is a range of standard deviation or variance through which we measure this thing.

Skewness is the descriptive major of the symmetry of this random variable with respect to again its central value, the coefficient of this Skewness is the measure of how we measure this one.

And next comes this kurtosis or tailedness, it provides the important information about the tails of this probability or how to pick the distribution is such as it is with respect to his outlier and all and we measure it with respect to the coefficient of kurtosis.

(Refer Slide Time: 32:41)



Measure of Central Tendency

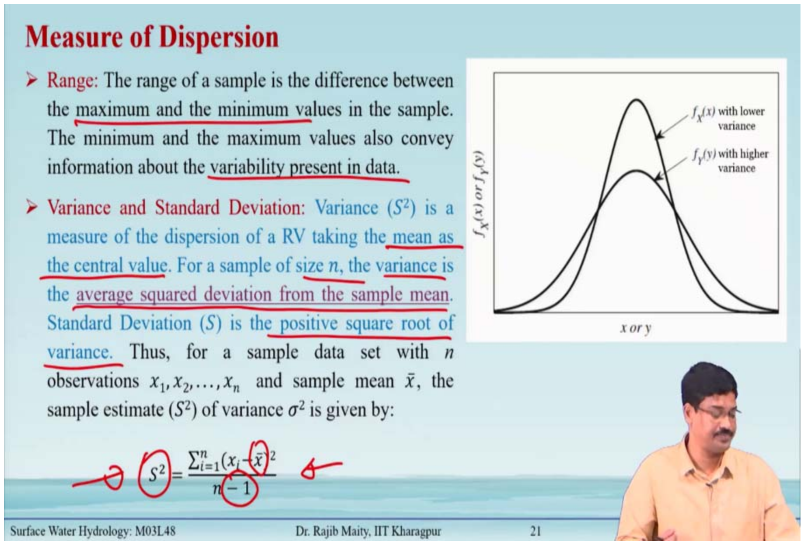
- Mean: Mean can be defined as the sum of the observations divided by sample size. Let us consider a sample data set with n observations $x_1, x_2, \ldots, x_n$ for a random variable$X$. The sample estimate ($\bar{x}$) of the mean is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Median: The median is the value of the RV at which the values on both sides of it are equally probable. Thus, it is the value of the RV with a 50% exceedance/non-exceedance probability.

- Mode: The mode is the most probable or most frequently occurring value of a random variable.

(Refer Slide Time: 34:14)



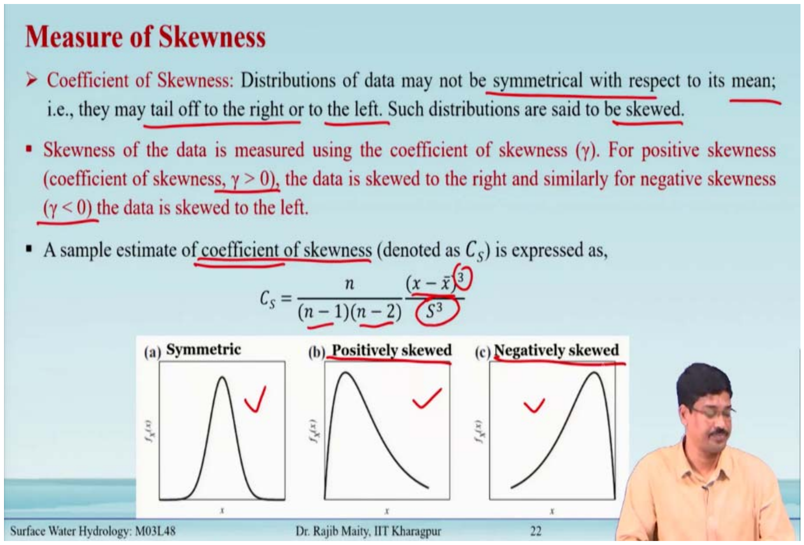
Measure of Dispersion

- Range: The range of a sample is the difference between the maximum and the minimum values in the sample. The minimum and the maximum values also convey information about the variability present in the data.

- Variance and Standard Deviation: Variance (S^2) is a measure of the dispersion of an RV taking the mean as the central value. For a sample of size$n$, the variance is the average squared deviation from the sample mean. Standard Deviation (S) is the positive square root of variance. Thus, for a sample data set with n observations $x_1, x_2, \ldots, x_n$ and sample mean $\bar{x}$, the sample estimate (S^2) of variance σ^2 is given by:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

(Refer Slide Time: 35:26)



Measure of Skewness

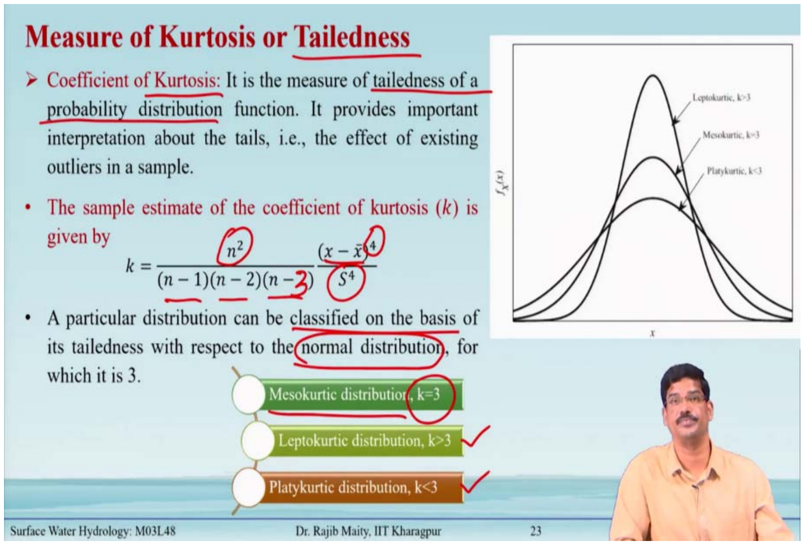
- Coefficient of Skewness: Distributions of data may not be symmetrical with respect to their mean; i.e., they may be tail off to the right or the left. Such distributions are said to be skewed.

- The Skewness of the data is measured using the coefficient of Skewness (γ). For positive Skewness (coefficient of Skewness, γ > 0), the data is skewed to the right, and similarly for negative Skewness (γ < 0) the data is skewed to the left.

- A sample estimate of the coefficient of Skewness (denoted as C_S) is expressed as,

$$C_S = \frac{n}{(n-1)(n-2)} \frac{(x-\bar{x})^3}{S^3}$$

(Refer Slide Time: 36:36)



Measure of Kurtosis or Tailedness

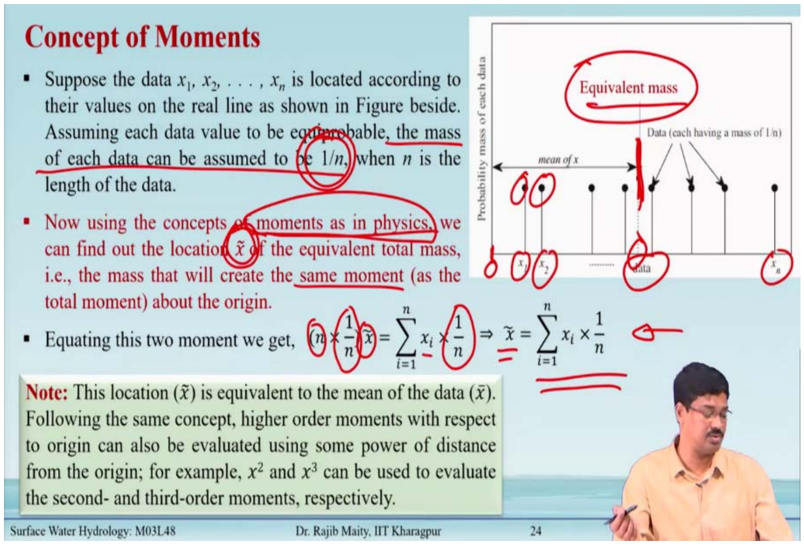
➢ Coefficient of Kurtosis: It is the measure of tailedness of a probability distribution function. It provides important interpretation about the tails, i.e., the effect of existing outliers in a sample.

The sample estimate of the coefficient of kurtosis (k) is given by

$$k = \frac{n^2}{(n-1)(n-2)(n-2)} \frac{(x-\bar{x})^4}{S^4}$$

A particular distribution can be classified based on its tailedness with respect to the normal distribution, for which it is 3. So, far as respect to the kurtosis, the particular distribution can be classified with respect to its tailedness and it is called the mesokurtic, if the k value is equal to 3 and the k value is equal to 3 for normal distribution, and it is leptokurtic if it is greater than 3 and platykurtic if it is less than 3.

(Refer Slide Time: 38:38)



Concept of Moments

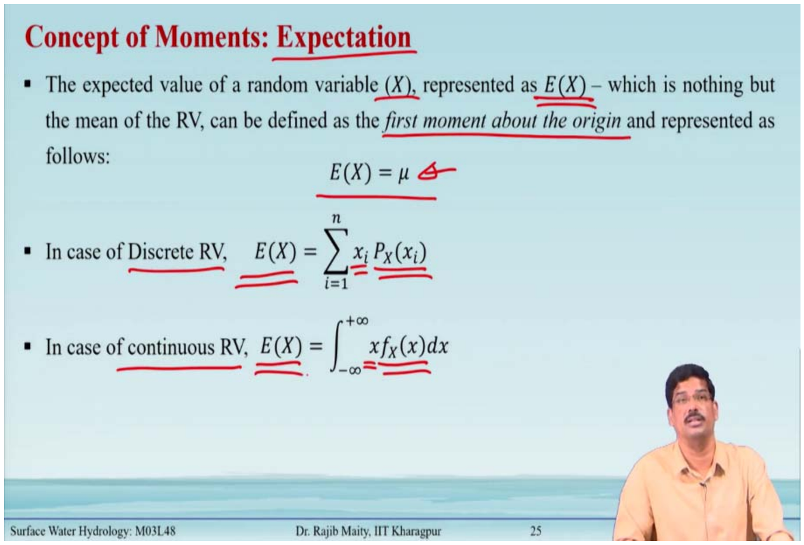
- Suppose the data $x_1, x_2 \ldots x_n$ is located according to their values on the real line as shown in Figure beside. Assuming each data value to be equiprobable, the mass of each data can be assumed to be $1/n$, when n is the length of the data.

- Now using the concepts of moments as in physics, we can find out the location $\tilde{x}$ of the equivalent total mass, i.e., the mass that will create the same moment (as the total moment) about the origin.

- Equating these two moments we get, $(n \times \frac{1}{n})\tilde{x} = \sum_{i=1}^{n} x_i \times \frac{1}{n} \Rightarrow \tilde{x} = \sum_{i=1}^{n} x_i \times \frac{1}{n}$

It may be noted that this location ($\tilde{x}$) is equivalent to the mean of the data ($\bar{x}$). Following the same concept, higher-order moments with respect to origin can also be evaluated using some power of distance from the origin; for example, x^2 and x^3 can be used to evaluate the second-and third-order moments, respectively.

(Refer Slide Time: 41:09)



Concept of Moments: Expectation

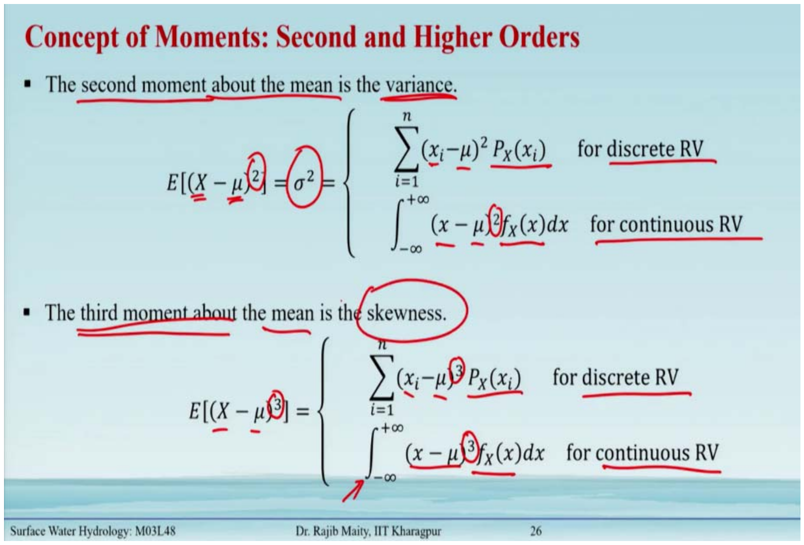
- The expected value of a random variable (X), represented as $E(X)$ – which is nothing but the mean of the RV, can be defined as the first moment about the origin and represented as follows:

$$E(X) = \mu$$

- In the case of Discrete RV, $E(X) = \sum_{i=1}^{n} x_i \, P_X(x_i)$

- In the case of continuous RV, $E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$

(Refer Slide Time: 41:58)



Concept of Moments: Second and Higher Orders

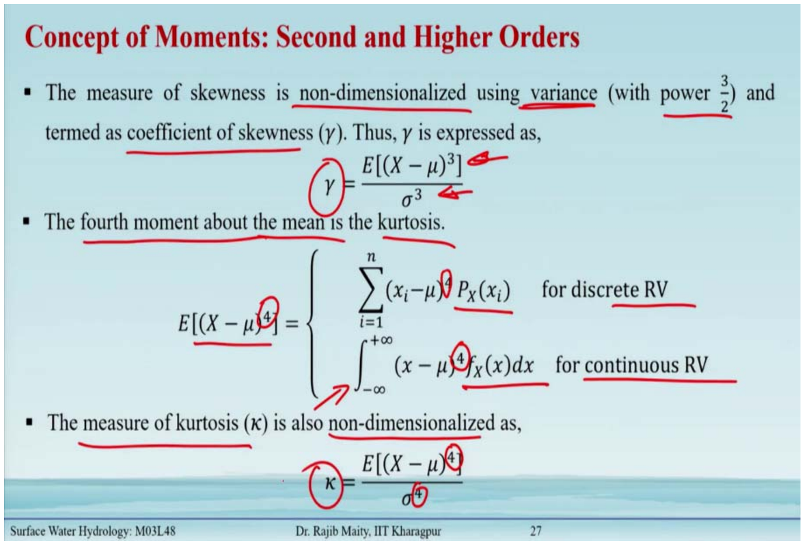
The second moment about the mean is the variance.

$$E[(X-\mu)^2] = \sigma^2 = \begin{cases} \sum_{i=1}^{n} (x_i-\mu)^2 \, P_X(x_i) & \text{for discrete RV} \\ \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)dx & \text{for continuous RV} \end{cases}$$

The third moment about the mean is the Skewness.

$$E[(X-\mu)^3] = \begin{cases} \sum_{i=1}^{n} (x_i-\mu)^3 \, P_X(x_i) & \text{for discrete RV} \\ \int_{-\infty}^{+\infty} (x-\mu)^3 f_X(x)dx & \text{for continuous RV} \end{cases}$$

(Refer Slide Time: 43:18)



The measure of Skewness is non-dimensionalized using variance (with power $\frac{3}{2}$) and termed as the coefficient of Skewness (γ). Thus, γ is expressed as,

$$\gamma = \frac{E[(X-\mu)^3]}{\sigma^3}$$

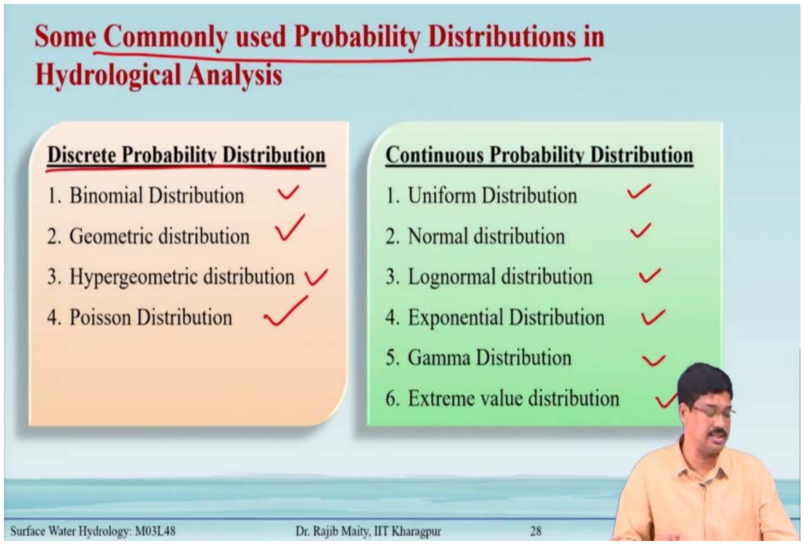
The fourth moment about the mean is the kurtosis.

$$E[(X-\mu)^4] = \begin{cases} \sum_{i=1}^{n}(x_i-\mu)^4 P_X(x_i) & \text{for discrete RV} \\ \int_{-\infty}^{+\infty}(x-\mu)^4 f_X(x)dx & \text{for continuous RV} \end{cases}$$

- The measure of kurtosis (κ) is also non-dimensionalized as,

$$\kappa = \frac{E[(X-\mu)^4]}{\sigma^4}$$

(Refer Slide Time: 44:18)



Some commonly used Probability Distributions in Hydrological Analysis

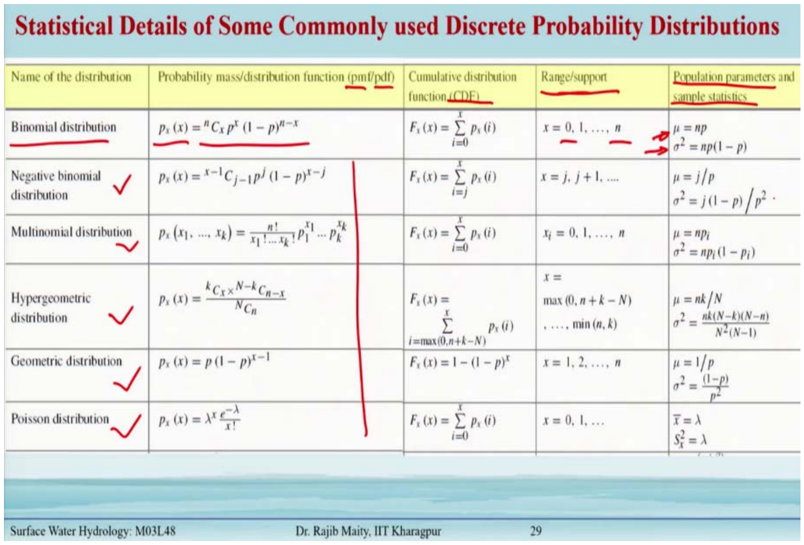
Discrete Probability Distribution

1. Binomial Distribution
2. Geometric distribution
3. Hypergeometric distribution
4. Poisson Distribution

Continuous Probability Distribution

1. Uniform Distribution
2. Normal distribution
3. Lognormal distribution
4. Exponential Distribution
5. Gamma Distribution
6. Extreme value distribution

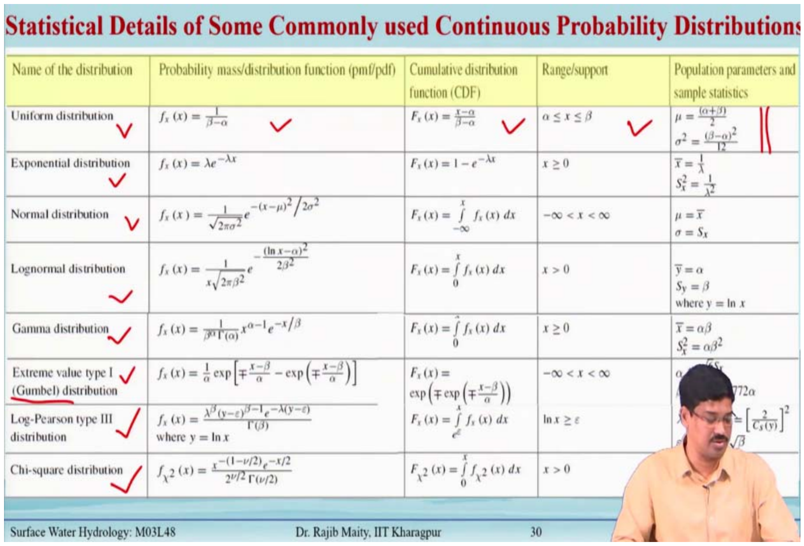
(Refer Slide Time: 45:01)



Statistical Details of Some Commonly used Discrete Probability Distributions

Name of the distribution	Probability mass/distribution function (pmf/pdf)	Cumulative distribution function (CDF)	Range/support	Population parameters and sample statistics
Binomial distribution	$p_x(x) = {}^nC_x p^x (1-p)^{n-x}$	$F_x(x) = \sum_{i=0}^{x} p_x(i)$	$x = 0, 1, \ldots, n$	$\mu = np$ $\sigma^2 = np(1-p)$
Negative binomial distribution	$p_x(x) = {}^{x-1}C_{j-1} p^j (1-p)^{x-j}$	$F_x(x) = \sum_{i=j}^{x} p_x(i)$	$x = j, j+1, \ldots$	$\mu = j/p$ $\sigma^2 = j(1-p)/p^2$
Multinomial distribution	$p_x(x_1, \ldots, x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$	$F_x(x) = \sum_{i=0}^{x} p_x(i)$	$x_i = 0, 1, \ldots, n$	$\mu = np_i$ $\sigma^2 = np_i(1-p_i)$
Hypergeometric distribution	$p_x(x) = \frac{{}^kC_x \times {}^{N-k}C_{n-x}}{{}^NC_n}$	$F_x(x) = \sum_{i=\max(0,n+k-N)}^{x} p_x(i)$	$x = \max(0, n+k-N), \ldots, \min(n, k)$	$\mu = nk/N$ $\sigma^2 = \frac{nk(N-k)(N-n)}{N^2(N-1)}$
Geometric distribution	$p_x(x) = p(1-p)^{x-1}$	$F_x(x) = 1-(1-p)^x$	$x = 1, 2, \ldots, n$	$\mu = 1/p$ $\sigma^2 = \frac{(1-p)}{p^2}$
Poisson distribution	$p_x(x) = \lambda^x \frac{e^{-\lambda}}{x!}$	$F_x(x) = \sum_{i=0}^{x} p_x(i)$	$x = 0, 1, \ldots$	$\overline{x} = \lambda$ $S_x^2 = \lambda$

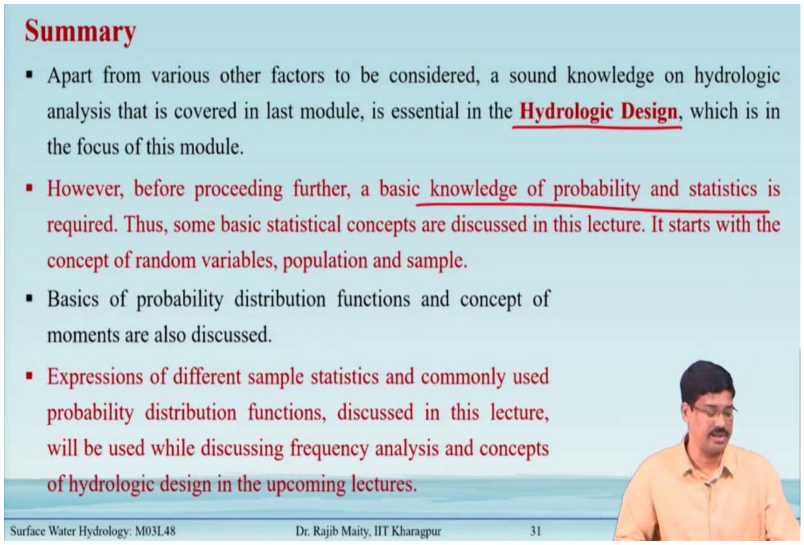
(Refer Slide Time: 46:22)



Statistical Details of Some Commonly used Continuous Probability Distributions

Name of the distribution	Probability mass/distribution function (pmf/pdf)	Cumulative distribution function (CDF)	Range/support	Population parameters and sample statistics
Uniform distribution	$f_x(x) = \frac{1}{\beta-\alpha}$	$F_x(x) = \frac{x-\alpha}{\beta-\alpha}$	$\alpha \le x \le \beta$	$\mu = \frac{(\alpha+\beta)}{2}$ $\sigma^2 = \frac{(\beta-\alpha)^2}{12}$
Exponential distribution	$f_x(x) = \lambda e^{-\lambda x}$	$F_x(x) = 1 - e^{-\lambda x}$	$x \ge 0$	$\overline{x} = \frac{1}{\lambda}$ $S_x^2 = \frac{1}{\lambda^2}$
Normal distribution	$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$	$F_x(x) = \int_{-\infty}^{x} f_x(x)\, dx$	$-\infty < x < \infty$	$\mu = \overline{x}$ $\sigma = S_x$
Lognormal distribution	$f_x(x) = \frac{1}{x\sqrt{2\pi\beta^2}} e^{-\frac{(\ln x-\alpha)^2}{2\beta^2}}$	$F_x(x) = \int_{0}^{x} f_x(x)\, dx$	$x > 0$	$\overline{y} = \alpha$ $S_y = \beta$ where $y = \ln x$
Gamma distribution	$f_x(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$	$F_x(x) = \int_{0}^{x} f_x(x)\, dx$	$x \ge 0$	$\overline{x} = \alpha\beta$ $S_x^2 = \alpha\beta^2$
Extreme value type I (Gumbel) distribution	$f_x(x) = \frac{1}{\alpha} \exp\left[\mp \frac{x-\beta}{\alpha} - \exp\left(\mp \frac{x-\beta}{\alpha}\right)\right]$	$F_x(x) =$ $\exp\left(\mp \exp\left(\mp \frac{x-\beta}{\alpha}\right)\right)$	$-\infty < x < \infty$	$\alpha = \frac{\sqrt{6}S_x}{\pi}$ $\beta = \overline{x} \mp 0.5772\alpha$
Log-Pearson type III distribution	$f_x(x) = \frac{\lambda^\beta (y-\varepsilon)^{\beta-1} e^{-\lambda(y-\varepsilon)}}{\Gamma(\beta)}$ where $y = \ln x$	$F_x(x) = \int_{e^\varepsilon}^{x} f_x(x)\, dx$	$\ln x \ge \varepsilon$	$\lambda = \frac{S_y}{\sqrt{\beta}}$, $\beta = \left[\frac{2}{C_s(y)}\right]^2$ $\varepsilon = \overline{y} - S_y\sqrt{\beta}$
Chi-square distribution	$f_{\chi^2}(x) = \frac{x^{-(1-\nu/2)} e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$	$F_{\chi^2}(x) = \int_{0}^{x} f_{\chi^2}(x)\, dx$	$x > 0$	$\mu = \nu$ $\sigma^2 = \nu^2$

(Refer Slide Time: 47:21)



Summary

In summary, we learned the following points from this lecture:

- Apart from various other factors to be considered, a sound knowledge of hydrologic analysis that is covered in the last module, is essential in the Hydrologic Design, which is the focus of this module.
- However, before proceeding further, a basic knowledge of probability and statistics is required. Thus, some basic statistical concepts are discussed in this lecture. It starts with the concept of random variables, population, and sample.
- Basics of probability distribution functions and the concept of moments are also discussed.
- Expressions of different sample statistics and commonly used probability distribution functions, discussed in this lecture, will be used while discussing frequency analysis and concepts of hydrologic design in the upcoming lectures.