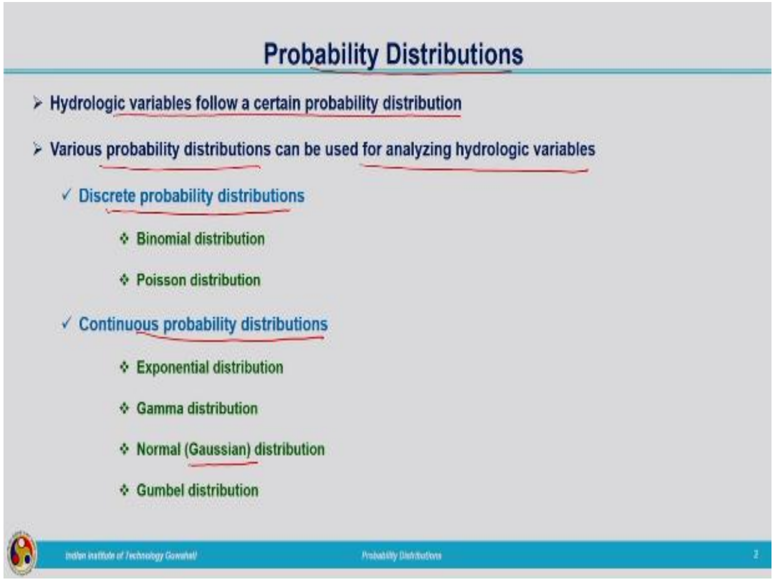
Engineering Hydrology
Dr. Sreeja Pekkat
Department of Civil Engineering
Indian Institute of Technology, Guwahati
Module: 6
Lecture 74: Probability Distributions

Hello all, welcome back. In the previous two lectures we were discussing about basics of probability and statistics which are required for the analysis of hydrologic variables. Hydrologic variables we have seen that it is involved with certain uncertainties, so in order to deal with the variables which are having uncertainties, we need to treat that in probabilistic way, that is why we need to have the basics related to probability and statistics. So, very minimal way we have covered in the previous two lectures, that is the different types of random variables, that is variables which are involved with uncertainties are considered as random variables and in that itself we have seen discrete random variables and continuous random variables.

After that we have seen what is meant by probability distribution, that is corresponding to each and every value taken up by the random variable, there is an associated probability. The relationship between these values and the probabilities is represented by means of probability distribution. In the case of continuous random variables, we will describe by means of probability density function and in the case of discrete random variables we make use of probability mass function. Then corresponding cumulative distribution functions also, we have looked into. After that we have discussed about the descriptive statistics which are required for hydrologic analysis, those were of measures of central tendency, measure of peakedness, measure of symmetry and measure of dispersion.

(Refer Slide Time: 3:16)

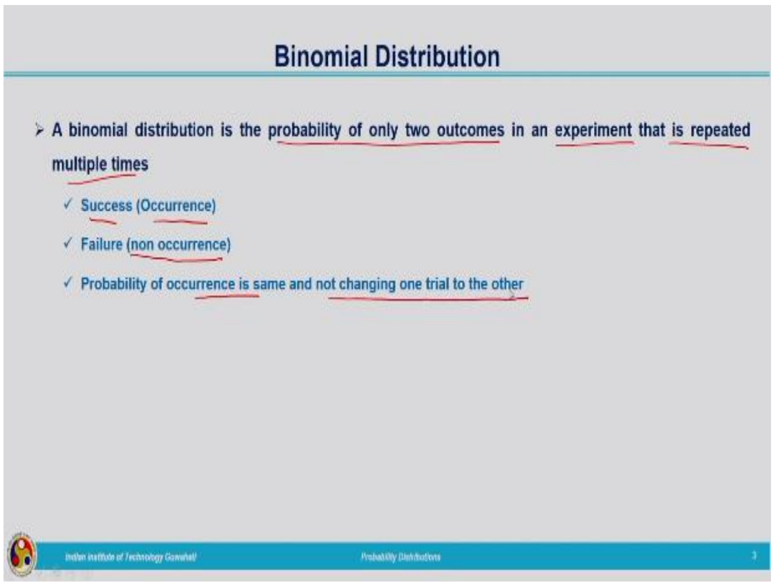


Now, let us look into different types of probability distribution functions which we commonly use in hydrology. Probability distribution function related to discrete random variable and also continuous random variable. So, we will start with probability distributions. Hydrologic variables follow certain probability distribution. Various probability distributions can be used for analyzing hydrologic variables. Various types of probability distribution functions are there but depending on the type of the variable which we are dealing with certain kind of probability distribution it will be following. So, majority of the probability distribution functions which we use in hydrologic analysis we will discuss here.

Discrete probability distributions and continuous probability distributions. Discrete probability distributions are related to discrete random variable and the continuous probability distributions are related to continuous random variables. In the case of number of rainy days, they will be taking integer values, those are considered as discrete random variables. But in the case of rainfall, stream flow etcetera, these will be taking a value within certain range, it cannot be expressed by means of an integer, it may be taking the fractional values also, that also within certain range. Those type of hydrologic variables will be considered as continuous random variables. The probability distributions which they follow are considered as the continuous probability distributions.

Under discrete probability distributions we will see binomial distribution and Poisson distribution and under continuous probability distribution we will look into commonly used exponential distribution, gamma distribution, normal distribution and Gumbel distribution, normal distribution is also termed as Gaussian distribution. So, these are the probability distributions which we are going to look into in this lecture.

(Refer Slide Time: 5:28)



<u>Binomial distribution</u>

A binomial distribution is the probability of only two outcomes in an experiment that is repeated multiple times. That is in this case the outcomes are only two. You can consider the case of tossing a coin, so what are the two different types of outcomes which you are expecting, one is getting a head and the second one is getting a tail. So, in both the cases the probability of occurrence is the same, that is 0.5. When you toss, the probability of getting a head is 0.5, the same probability is there for getting a tail also. That value of probability is not changing and at the same time we can expect only two outcomes; either a head or a tail. So, this is that kind of distribution in which we can expect only two outcomes for different number of trials.

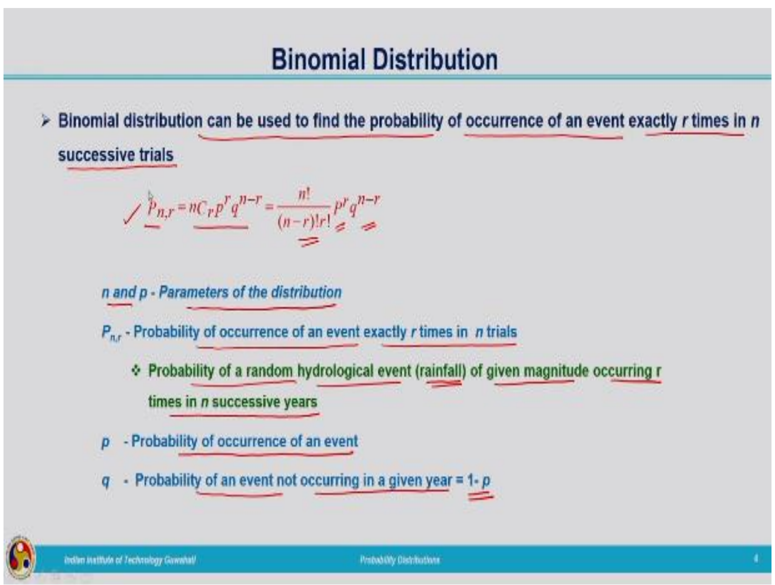
So, the outcomes can be termed as

- ✓ Success or occurrence
- ✓ Failure or non-occurrence

After tossing a coin, if probability of getting a head is considered as the success (or occurrence) means failure is probability of getting a tail. So, these outcomes are considered as success and failure, that is occurrence and non-occurrence.

In the case of a flood, whether there is a flood or there is no flood, these are two cases, yes or no cases, there can be a flood in this year, there cannot be a flood. So, we are not quantifying the values, probability of occurrence of a flood, whether there will be a flood in the coming year or there will not be any flood. So, that can be represented by means of two outcomes, that is one chance is that there can be a flood and on the other hand there cannot be a flood. So, these types of analysis can be done by making use of binomial distribution.

(Refer Slide Time: 8:05)



The probability of occurrence is same and not changing one trial to other. Binomial distribution can be used to find the probability of occurrence of an event exactly *r* times in *n* successive trials (*n* times the experiment is conducted, in that we are expecting the probability of occurrence of an event *r* times, exactly *r* times). Then we can represent that probability

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

- ✓ n and p - the parameters of the distribution,
- ✓ n - number of trials and
- ✓ p - probability of occurrence of an event
- ✓ $P_{n,r}$ - the probability of occurrence of an event exactly r times in n trials.

We are telling, probability of getting a head r times in n trials, if that is considered as success, then failure is probability of getting tail in n - r numbers out of n experimental trials.

So, we do not have to separately mention what is the probability of failure, it can be written as 1 minus probability of success. That is why total number of trials or total number of experiments 'n' and the probability of getting a particular event (r times) 'p' are considered as parameters of distribution.

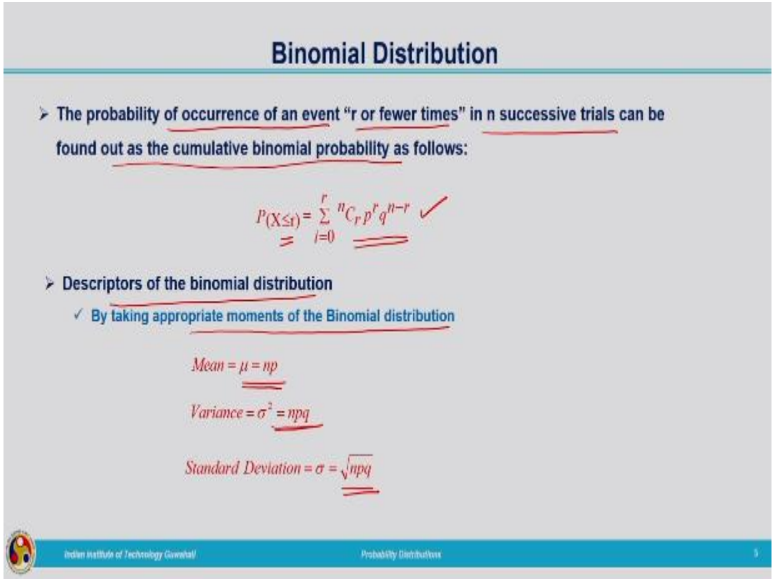
Now, $P_{n,r}$ is the probability of occurrence of an event exactly r times in n trials, that is probability of a random hydrological event. For example, in case of rainfall, it is probability of getting a rainfall of given magnitude occurring r times in n successive years. Sometimes based on the value of rainfall we can tell whether there will be a flood or not.

We commonly use a term related to probability of accidents, what is meant by probability of accidents? It is the probability of occurrence of an event for which the value is greater than certain set point value. We are having certain rainfall value, beyond that if rainfall is occurring we will be considering it as chances of flooding is there in that particular year.

So, that way by analyzing the rainfall data you can find out the probability of getting a rainfall which is more than that of the set point value. So, based on that we can tell there may be a flood or there may not be a flood. So, if beyond certain set point value it is coming, then that probability is termed as probability of accidents.

So, this $P_{n,r}$ is giving us the probability of accidents and p is the probability of occurrence of an event. Whether there is a success, that is represented by p. Then q is the probability of an event not occurring in the given year that can be represented by $1 - p$. This $P_{n,r}$ represents the binomial distribution.

(Refer Slide Time: 12:09)



Now, let us move on to the cumulative probability. The probability of occurrence of an event *r* or fewer times in *n* successive trials can be found as the cumulative binomial probability. That can be represented as

$$P_{(X \le r)} = \sum_{i=0}^{r} {}^{n}C_r \, p^r q^{n-r}$$

We are calculating the summation of the probability mass function. That is this binomial distribution is corresponding to discrete random variable, in that case we will be calling the relationship by means of probability mass function. Cumulative of that will be giving you the cumulative distribution function that is given by the above expression. So, this much about the binomial distribution and cumulative binomial distribution.

Now, let us move on to the descriptive statistics related to binomial distribution that is descriptors of the binomial distribution. By taking appropriate moments of the binomial distribution we can get the descriptive statistics related to a particular distribution. All these descriptors are calculated by considering the moments of the distribution, that is in the case of mean we have considered the first moment with respect to origin and other descriptors such as the variance, skewness and kurtosis, those are the moments taken with respect to mean value, not with respect to origin. So, the mean in the case of binomial distribution is given by

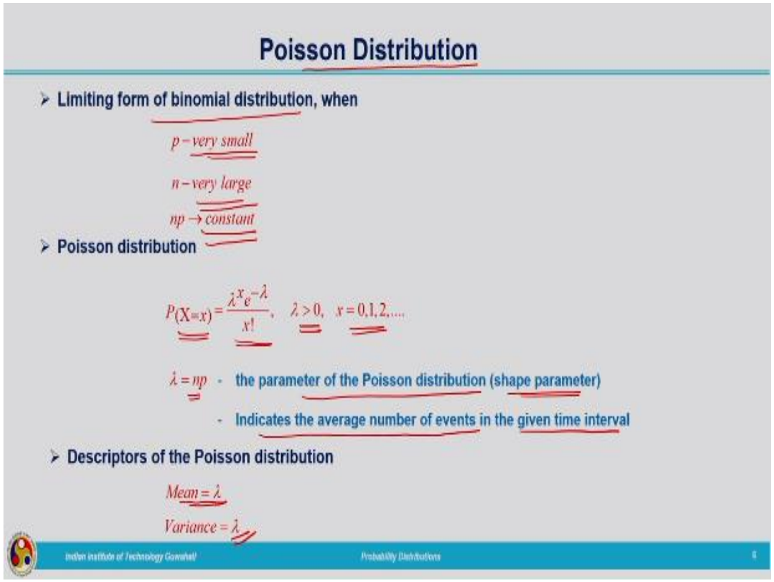
Mean $= \mu = np$

Variance $= \sigma^2 = npq$

standard deviation $= \sigma = \sqrt{npq}$

These are the descriptors of the binomial distribution.

(Refer Slide Time: 14:47)



Poisson distribution

Poisson distribution is the limiting form of binomial distribution. In this case we are considering probability of occurrence of a particular event p to be very small but the number of experiments or number of trials considered, n is very large.

So, we get $np \longrightarrow$ constant.

Poisson distribution is given by the function

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \lambda > 0, \quad x = 0,1,2,....$$

In this case $\lambda = np$, that is considered as the parameter of the Poisson distribution. This is named as shape parameter of the probability distribution. This indicates the average number of events in the given time interval. A certain time interval is considered, within that interval how much is the average number of events, that is considered as the value λ.

Now, let us move on to the descriptors of Poisson distribution.

Mean = λ

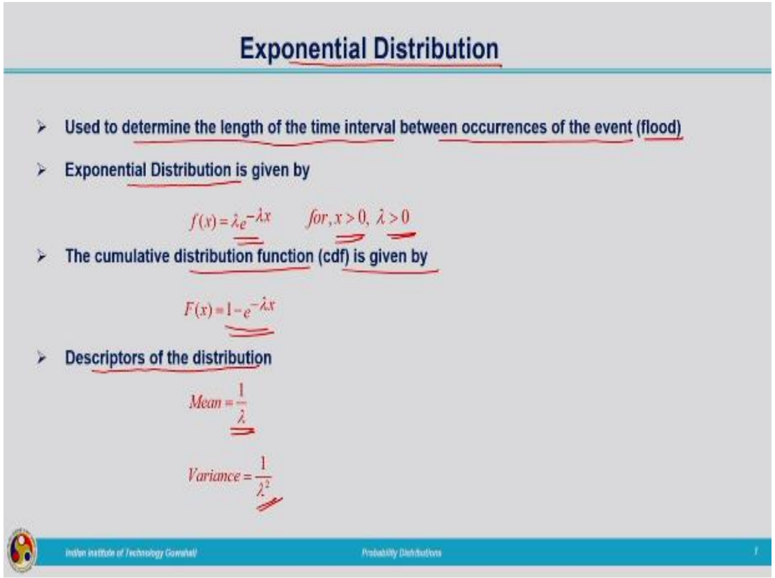
Variance = λ

standard deviation = $\sigma = \sqrt{npq}$

In the case of Poisson distribution, it is a limiting case of the binomial distribution. Different descriptors are there that is mean, variance, standard deviation, coefficient of skewness, coefficient of kurtosis, all those things are there but I am not going to higher moments, I am talking about first and second moments only, that is the mean and variance. In the case of Poisson distribution mean is λ, that is the parameter of the distribution and the same as that of the variance.

So, these are the two distributions related to discrete random variables which we use in hydrology. Some others are also there, that you study when you go in advanced level. So, here we have seen two discrete probability distribution functions that is binomial distribution and the Poisson distribution.

(Refer Slide Time: 18:01)



Now, let us move on to the continuous probability distribution functions. First one we are going to discuss is exponential distribution.

<u>Exponential distribution</u>

It is used to determine the length of the time interval between the occurrences of the events. For example, flood, that is it gives an idea about length of the time interval between occurrence of two flood events.

Exponential distribution is given by the function

$$f(x)=\lambda e^{-\lambda x} \quad for, x>0,\ \lambda>0$$

Since, this is for continuous random variable, we will be expressing by means of probability density function ($f(x)$). While talking about the values of λ and x, x can take any value > 0, λ can also take any value > 0.

When we talk about hydrologic variables majority of the hydrologic variables are positive, rainfall value, stream flow value, evaporation value all these values are positive. So, in the case of exponential distribution x and λ are > 0.

Now, coming to the cumulative distribution function of exponential distribution, that is if we know the probability density function, integral of that probability density function will be giving you the cumulative distribution function.

So, cdf is given by

$$F(x) = 1 - e^{-\lambda x}$$

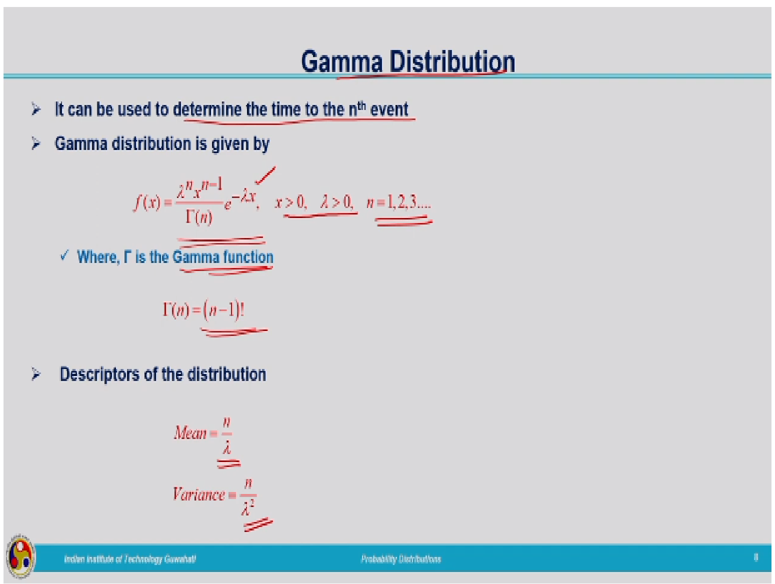
Now, next details which we need is related to descriptors of the distribution. Coming to the descriptors,

$$Mean = \frac{1}{\lambda}$$

$$Variance = \frac{1}{\lambda^2}$$

If variance is $1/\lambda^2$, standard deviation will be $1/\lambda$.

(Refer Slide Time: 20:44)



Now, coming to the next probability density function corresponding to continuous random variable that is gamma distribution.

Gamma distribution

It can be used to determine the time to the n^{th} event. So, many events are occurring independent events so time to the n^{th} event can be obtained by making use of gamma distribution. It is given by the function

$$f(x) = \frac{\lambda^n x^{n-1}}{\Gamma(n)} e^{-\lambda x}, \quad x > 0, \quad \lambda > 0, \quad n = 1, 2, 3....$$

Γ is gamma function, this we have seen while explaining Nash model, so,

$$\Gamma(n) = (n-1)!$$

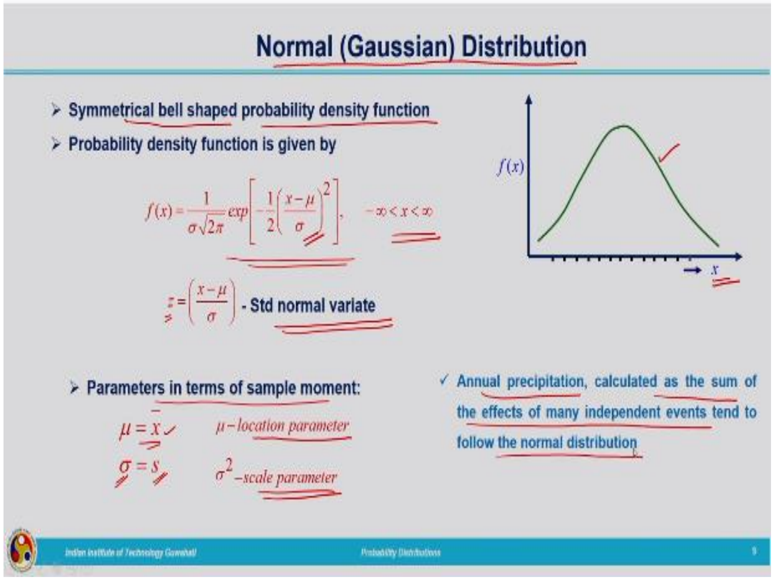
Now, coming to the descriptors of the gamma distribution,

$$Mean = \frac{n}{\lambda}$$

$$Variance = \frac{n}{\lambda^2}$$

In this case you can compare it with the exponential distribution, exponential distribution the descriptors were found to be mean is $1/\lambda$, variance is $1/\lambda^2$. Here, for each event equal weightage is given that is why mean is given by n/λ and variance is given by n/λ^2.

(Refer Slide Time: 22:44)



Now, we will move on to the next continuous probability distribution that is the one which is commonly used in hydrologic analysis which is termed as normal distribution and also Gaussian distribution.

<u>Gaussian distribution</u>

So, Gaussian distribution is very common in hydrologic analysis, it has got a symmetrical bell-shaped probability density function. Probability density function can be schematically represented by means of a bell-shaped function which can be schematized like this.

Our probability density function is along the y-axis and along the x-axis the values of x is given and if you plot the pdf, it will be looking like a bell-shaped function. This is very important function which we commonly use in the hydrologic analysis. As the number of years, we are considering the annual rainfall data and we are trying to fit some probability distribution for those data, in that case if the data is of annual data then we can make use of normal distribution but in other cases, that is if you are dealing with the hourly rainfall data or daily rainfall data we cannot make use of this particular distribution.

This is a symmetrical bell-shaped probability density function and the function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \qquad -\infty < x < \infty$$

The range of these distribution that is x is between - ∞ to + ∞. But majority of the hydrological variables are positive, rainfall either there would not be any rainfall 0 or after that it can take any value, any positive real number. So, the value will be between 0 and ∞ when we talk about hydrologic variables.

Now,

$$z = \left(\frac{x-\mu}{\sigma}\right)$$

z is termed as standard normal variate. We can write the probability density function in terms of standard normal variate also, that is in terms of z.

Now, coming to the parameters of the distribution that is

$$\mu = \bar{x}$$

μ - the population mean and $\bar{x}$ - the sample mean.

Population mean can be represented by means of sample mean and σ. That is

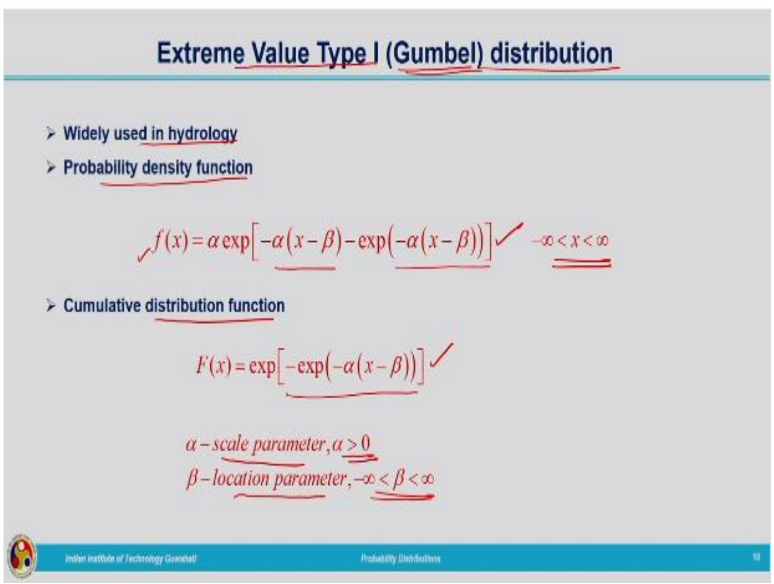
$$\sigma = s$$

σ is the standard deviation corresponding to population and s is the notation used for the standard deviation in the case of sample. So, in the case of normal distribution we can consider the population mean and population standard deviation as the sample mean and sample standard deviation.

In this case μ - location parameter and σ^2 – (variance) is the scale parameter, The μ is representing the average value and the variance σ^2 is representing the dispersion in the data. So, based on this mean value and the dispersion value, we can understand how the shape or the how the data is spread, the details related to data can be understood by going through the shape parameter and the location parameter.

So, that much about normal distribution, very minimal description I am giving over here. Now, for example, in the case of annual precipitation, we commonly make use of normal distribution as the probability distribution. Annual precipitation that is calculated as the sum of the effects of many independent events tend to follow normal distribution. Many independent events means each and every day's rainfall is independent of the previous day or next day's rainfall. So, annual precipitation is calculated as the sum of the daily precipitation. So, that usually follow the normal distribution but it is not the case with the daily precipitation or hourly precipitation, in that case we cannot make use of normal distribution.

(Refer Slide Time: 28:18)



Now, next one is the Extreme Value Type I distribution, which is termed as Gumbel's distribution.

Gumbel's distribution

When we talk about hydrologic analysis, extremes are too common nowadays, that is occurrence of a flood. During the monsoon season, majority of the places are getting flooded and urban cities are very commonly getting flooded, rivers are overflowing, flooding in that entire area is taking place, so extreme events are very common in monsoon season.

At the same time during non-monsoon season drought is also a common phenomenon. So, the scarcity of rain and also abundance of rain, so these two extremes are very common. For analyzing these type of hydrologic variables or hydrologic analysis involved with these extreme values we need to go for extreme value analysis. For that we commonly make use of extreme value type I distribution (Gumbel distribution), extreme value type II, type III distributions are there, so those I am not covering over here, I am going to discuss about only the extreme value type I distribution that is Gumbel's distribution.

It is widely used in hydrology, very commonly used in hydrology and the probability density function is given by

$$f(x) = \alpha \exp\left[-\alpha\left(x-\beta\right) - \exp\left(-\alpha\left(x-\beta\right)\right)\right]$$

So, this is the probability density function corresponding to Gumbel distribution. This is very commonly used in hydrologic analysis, especially in the case of frequency analysis. x can take the value between - ∞ to + ∞.

And cumulative distribution is given by

$$F(x) = \exp\left[-\exp\left(-\alpha\left(x-\beta\right)\right)\right]$$

In this α is the scale parameter, it can take value > 0 and β is the location parameter, it can take values between - ∞ to + ∞.

The scale parameter and location parameters are described by α and β in the case of Gumbel's distribution and we have looked into the probability density function and also cumulative distribution function in the case of extreme value type I Gumbel distribution. And this we will be making use and we will be studying more into this type I distribution when we discuss about frequency analysis. So, these are the very commonly used distributions in hydrology as far as discrete random variables and the continuous random variables are concerned. So, very minimal way I have covered in this lecture.

(Refer Slide Time: 31:38)

You can go through these references for getting more understanding. So, here I am winding up this lecture, thank you very much.