

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

Lecture 73: Probability Distribution and Basic Descriptive Statistics

Hello all, welcome back, in the previous lecture we have started with the module on Hydrologic Statistics. We have discussed about the role of probability and statistics in the hydrologic analysis and we have discussed about the basics related to probability and also different laws concerning the basic concepts of probability.

Today we will move on to the concepts related to probability distribution and some more about descriptive statistics. We have already seen that hydrologic variables are having certain uncertainties involved with that, so those type of variables can be expressed as random variables. Different types of random variables, discrete and continuous we have discussed in the previous lecture.

(Refer Slide Time: 1:47)

Probability Distribution

- Random variables can take any value
- Each value is associated with a probability
- For all values of the given random variable, corresponding probabilities are calculated
- Probability distribution
 - ✓ Relationship between the values of random variable and their corresponding probability values

Indian Institute of Technology Guwahati | Probability Distribution and Basic Descriptive Statistics

Today let us start with the probability distribution. When we talk about random variables, it can take any value within certain range related to that particular variable. So, random variables can take any value and each value is associated with a probability. For the value assigned to a random variable corresponding to that there will be a probability, so that way different values

can be taken up by random variables and for each of these values, there will be corresponding probabilities associated with that.

For all values of the given random variable, corresponding probabilities can be calculated and the relationship between the values of random variable and their corresponding probability values is discussed by means of probability distribution. What is meant by probability distribution? For example, if you are considering a variable that is having a fixed value, then there would not be any uncertainties involved with that, corresponding to that particular value as input we will be having single output, but that is not the case with the variables which are involved with uncertainties.

So, those variables, for a single value we will be having different outputs at the same time at the same location, depending on the uncertainties involved with that. So, each and every value which are taken up by the random variable can be associated with certain probability. So, if you are finding out a relationship with the value corresponding to the random variable and the corresponding probability value that is termed as the probability distribution.

(Refer Slide Time: 4:03)

Probability - Discrete Random Variable

➤ Probability mass function ✓

✓ Probability of discrete random variable

$$P(X = x) = p(x)$$

X - Random variable
x - specified value
P - probability that the RV X is equal to x

p(x) - Probability mass function

$$0 \leq p(x) \leq 1$$

✓ Using the law of total probability


$$\sum_{i=1}^N p(x_i) = 1$$

➤ Cumulative distribution function

✓ Cumulative probability of a discrete RV

$$P(X \leq x_i) = F(x_i) = \sum_{j=1}^{x_i} p(x_j)$$

✓ Useful for getting the probability of a discrete RV X having a value less than or equal to a specified value x_i


Indian Institute of Technology Guwahati
Probability Distribution and Basic Descriptive Statistics
2

Now, let us discuss about the probability distribution corresponding to discrete and continuous random variables. First, we will start with the discrete random variable. In the case of discrete

random variable, we will call this relationship as probability mass function. Probability mass function represents the probability of discrete random variable. It is represented by

$$P(X \leq x) = p(x)$$

In this expression X is the random variable, x is the specified value which can be taken up by the random variable X , P represents the probability that the random variable $X = x$. P is representing the probability corresponding to that specified value and $p(x)$ is representing the probability mass function, that is relating the value corresponding to the random variable and the respective probability related to that particular value of the random variable that is represented by means of probability mass function in the case of a discrete random variable.

$$0 \leq p(x) \leq 1$$

Now, using the law of total probability

$$\sum_{i=1}^N p(x_i) = 1$$

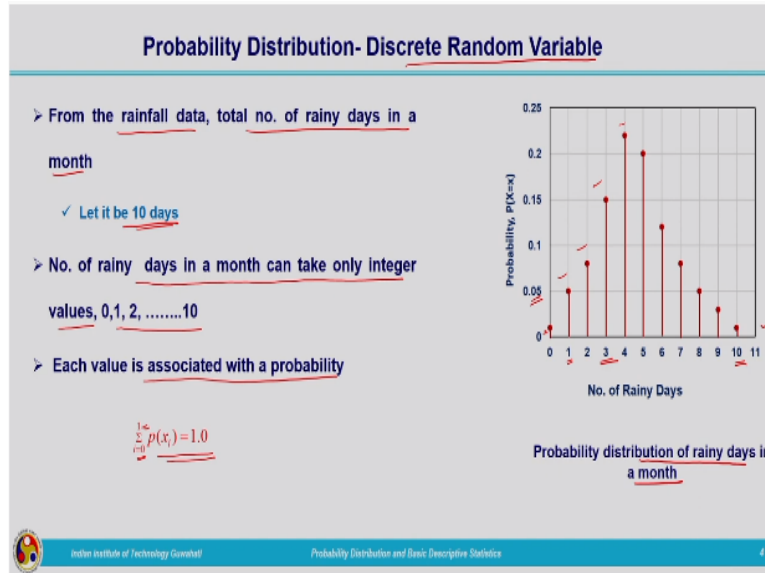
The next way of representation of discrete random variable is cumulative distribution function. As the name indicates, it is the cumulative of the probability corresponding to a value or less than that particular value. So, cumulative probability of a discrete random variable can be written as

$$P(X \leq x_i) = F(x_i) = \sum_{j=1}^i p(x_j)$$

This is useful for getting the probability of a discrete random variable X having a value \leq to a specified value x_i .

So, when we talk about the probability distribution related to discrete random variable, we express either in terms of probability mass function or cumulative distribution function.

(Refer Slide Time: 7:45)



Now, let us explain this with the help of an example. Consider the case of the discrete random variable, the probability distribution involves discrete random variables. If you are talking about hydrologic analysis, one example is the total number of rainy days in a particular month, that can be expressed in terms of certain integer values, those type of values are termed as the discrete random variables.

So, we are going to consider an example with total number of rainy days in a month. Consider from the rainfall data total number of rainy days in a month (you can consider the case of the month of June). We are having total 30 days and out of that, for example, let 10 days we are getting rain during the month of June. Number of rainy days in a month can take only integer values and these 10 days can take from 0 to 10, 0 represents no rainy day, 1 rainy day, 2 rainy days, that way up to 10 rainy days out of 30 days.

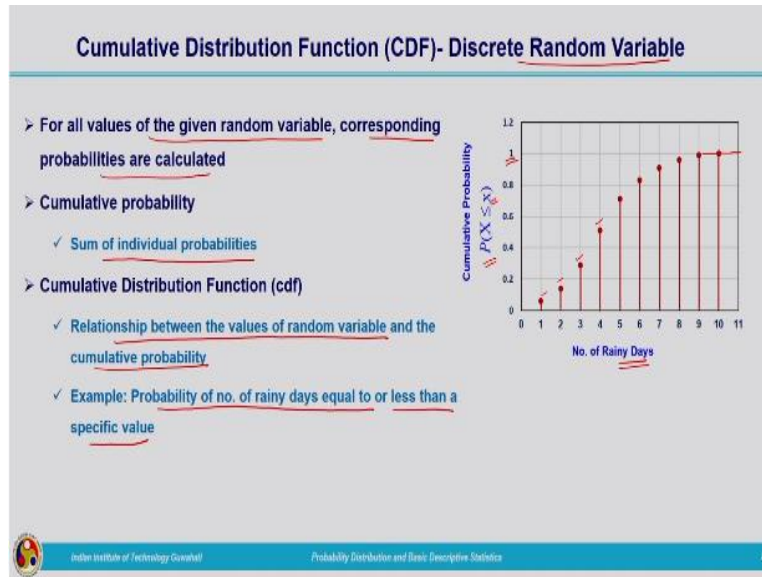
Each value is associated with a probability, we can assign certain probability for each and every value corresponding to the number of rainy days. The number of rainy days equal to 2, corresponding to that we can associate certain probability, so that way each and every value, that is the value which can be taken up by the random variable can be associated with certain probability.

$$\sum_{i=0}^1 p(x_i) = 1.0$$

So, this can be represented by means of a graph that is plotted here, that is the probability distribution of rainy days in a month against the number of rainy days. That is, we have considered a particular month, in that 10 days we are getting as rainy days and remaining 20 days we are getting as non-rainy days, so we can calculate the probability associated with each value corresponding to rainy day and plot in the graph. That is number of rainy days on the x-axis and probability along the y-axis, it is plotted as shown in this graph. Since, this is a discrete random variable, we are representing it by means of discrete representation. From this we can find out the probability associated with the number of rainy days equal to 3 or 4.

Number of rainy days can take only integer values, it cannot take any fraction value, either it is raining for 1 hour or 2 hours, depending on the intensity of rainfall we will be considering it as rainy day or if it is less than certain value, it will be considered as non-rainy day. So, that classification we have already seen while discussing about the classification of rainfall. So, based on that we can find out how many days are rainy days and how many days are non-rainy days and we can find out the associated probability that is plotted here that is representing the probability distribution of rainy days in a month.

(Refer Slide Time: 12:21)



Now, coming to cumulative distribution function of discrete random variable. Cumulative distribution function also we are making use of the same example which is considered for describing the probability mass function related to discrete random variable, that is the total number of rainy days.

For all values of the given random variable, corresponding probabilities are calculated as we have done in the previous slide. For rainy day is equal to 1, associated probability is calculated, that way up to 10 days, associated probabilities are calculated. After that we will be calculating the cumulative probability, that is sum of individual probabilities and that is represented by means of cumulative distribution function CDF.

CDF is nothing but the relationship between the values of random variable and the cumulative probability. In the previous case we were finding out the relationship with the random variable and the corresponding associated probability, but in the case of cumulative distribution function what we are doing, we are finding out the relationship with the random variable and the corresponding cumulative probability rather than individual probability.

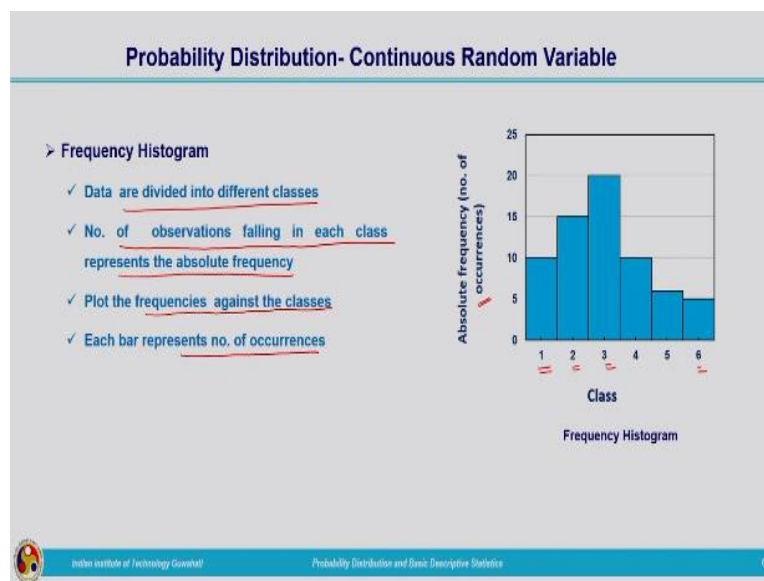
Example is probability of number of rainy days equal to or less than a specific value. We are not telling corresponding to rainfall equal to 2 days, we are not talking about the case as explained in

the previous slide, in that case we were relating the probability associated with the random variable corresponding to a specified value and the associated probability.

Here, instead of that we are talking about the number of rainy days ≤ 2 or number of rainy days ≤ 5 . That way we will get the representation based on cumulative distribution function. So, it can be plotted like this as in the previous example we have calculated the probabilities associated with each and every rainy day, that is added up to get the cumulative probability.

So, that is what is plotted over here, number of rainy days along the x-axis and cumulative probability $P(X) \leq x$ is plotted along the y-axis. It is an increasing curve, the value is increasing and finally it is reaching a value which is equal to 1, that is the total probability. So, cumulative values are taken up that is each and every probability is added up until it reaches the last value. Corresponding to the final value it will be equal to the total probability that is equal to 1. That much about the probability distribution about discrete random variable.

(Refer Slide Time: 15:43)



Now, let us look at the case with continuous random variable. In the case of continuous random variable, we cannot represent the variable with a single value, we will be representing the data or the random variable within a certain range. For example, if you are talking about the amount of rainfall which will be occurring tomorrow, that can be expressed within a range, we cannot represent it by means of a single value.

So, in this case we will be dividing the data into number of classes, that we can understand with the help of a frequency histogram. What is frequency histogram? The data corresponding to a particular variable are divided into different classes and the number of observations falling in each class represents the absolute frequency. After that we will plot the frequencies against the classes, each bar represents number of occurrences. So, that can be explained with the help of the frequency histogram. We are dividing the data into different classes, so number of occurrences will be calculated, that is represented by absolute frequency or number of occurrences. That absolute frequency versus number of classes will be plotted which will be providing us the frequency histogram.

(Refer Slide Time: 17:29)

Probability Distribution- Continuous random variable

➤ **Probability density function (pdf)**

✓ Probability of occurrence of a continuous RV is specified within an interval

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$


$f(x)$ - Probability density function

➤ **Cumulative distribution function (cdf)**

✓ Cumulative distribution represents the probability that X is less than or equal to a specific value, x

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx$$

$$f(x) = \frac{dF(x)}{dx}$$


Indian Institute of Technology Guwahati
Probability Distribution and Basic Descriptive Statistics

Now, coming to the probability distribution of continuous random variable. In the case of continuous random variable, we will be making use of probability density function PDF. What we have observed in the case of discrete random variable, it was represented by means of probability mass function, here it is density function, that is the probability of occurrence of a continuous random variable is specified within an interval. So, the value of x is specified within a range or interval like this,

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$

In this $f(x)$ is representing the probability density function.

So, you look at the left-hand side, here we are having the random variable X , it is taking a value between the range or between an interval x_1 to x_2 , it is not represented by means of a discrete value. That value can be calculated by finding out the interval of the probability density function

within that range, x_1 and x_2 , that is equal to $\int_{x_1}^{x_2} f(x) dx$.

Now, coming to cumulative distribution function. This cumulative distribution represents the probability that the random variable X is \leq to a specified value x , it can be represented like this,

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx$$

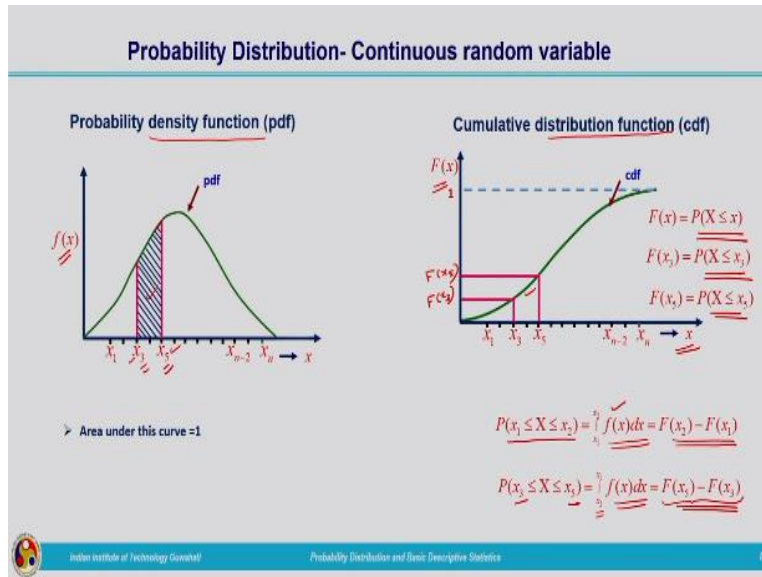
In the previous case we were taking summation, sigma. Here in this case this is applicable to the continuous random variable, so we are making use of the integral because it is within a range.

So, cumulative distribution represents the probability of a random variable $X \leq$ specified value.

Now, the relationship between cumulative distribution function and the probability density function is here, that is cumulative distribution function can be obtained by integrating the probability density function. So, we can find out the PDF, probability density function by differentiating the cumulative distribution function. So,

$$f(x) = \frac{dF(x)}{dx}$$

(Refer Slide Time: 20:49)



Now, we can plot the curves related to probability density function and cumulative distribution function in the case of a continuous random variable. So, probability density function PDF we will start, that is along the x-axis we are plotting the values which will be taken up by the random variable and along the y-axis the probability density function, corresponding to x_1, x_2 what is $f(x_1), f(x_2)$ that will be plotted along the y-axis and we can get a continuous curve.

So, this is our PDF, probability density function and the probability in the case of a continuous random variable, we are representing between certain interval, x_1 to x_2 or that interval will be specified, it is not same as that of the discrete random variable. So, here it is expressed in such a way that it is within x_3 and x_5 . So, the area marked by this blue hatched area is representing the probability corresponding to that particular range x_3 and x_5 and the total probability will be equal to 1, that is the area under this curve will be equal to 1.

Now, coming to the cumulative distribution function CDF, it is nothing but the cumulative of the probabilities which we have seen in the case of probability density function. So, that can be plotted like this. Along the x-axis again the values which are taken up by the random variable and along the y-axis we are having the cumulative probability. So, $f(x)$ can be calculated and plotted against x and it will be represented by a continuous curve as shown in this figure, this is our cumulative distribution function and the maximum value attained by this cumulative

distribution function will be equal to 1. So, that maximum value can be observed from this curve that is equal to 1.

Now, this expression corresponding to $F(x)$ is

$$F(x) = P(X \leq x)$$

x is the specified value corresponding to the random variable and if you are talking about the value corresponding to x_3 and x_5 , we can write

$$F(x_3) = P(X \leq x_3)$$

that can be marked in the curve (refer figure in slide).

Now, similar to this we can get the value corresponding to $F(x_5)$ given by

$$F(x_5) = P(X \leq x_5)$$

that can be marked here in the figure (refer slide).

Now, if we consider the probability corresponding to x within an interval of x_3 to x_5 , how can it be obtained? If we take the difference of this cumulative probability corresponding to x_5 and x_3 , it will be giving us the probability corresponding to the random variable within that interval. So, that can be written mathematically like this

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1)$$

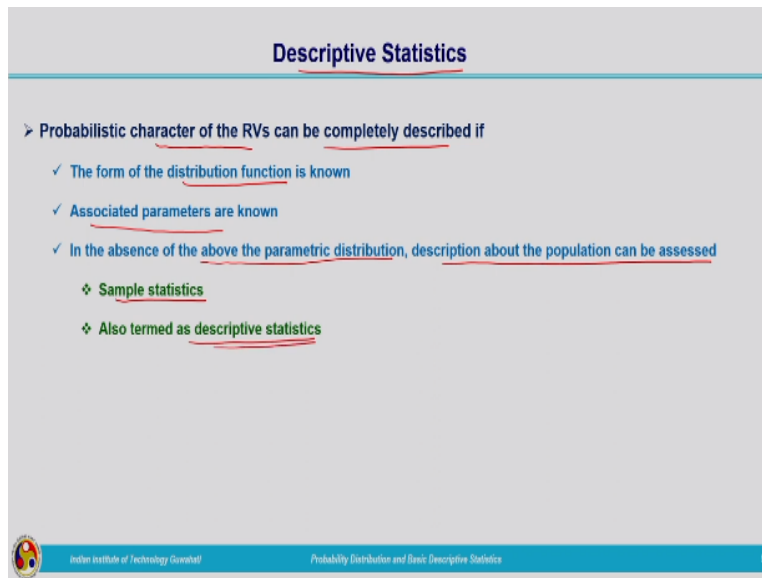
So, this is nothing but the difference in the cumulative distribution function corresponding to these two values x_1 and x_2 . So, corresponding to x_3 and x_5 we can write this expression,

$$P(x_3 \leq X \leq x_5) = \int_{x_3}^{x_5} f(x)dx = F(x_5) - F(x_3)$$

So, the difference between these cumulative probabilities corresponding to x_5 and x_3 will give us the probability corresponding to that particular random variable within that range. So, that much about probability density function and cumulative distribution function. Now, we need to have

understanding about different probability distribution functions. That we will see later, before that we need to have some basic understanding about the statistics.

(Refer Slide Time: 26:06)



The slide is titled "Descriptive Statistics" in a blue header. Below the title, there is a main bullet point: "Probabilistic character of the RVs can be completely described if". This is followed by three sub-bullet points, each starting with a checkmark: "The form of the distribution function is known", "Associated parameters are known", and "In the absence of the above the parametric distribution, description about the population can be assessed". Under the last sub-bullet point, there are two diamond-shaped icons: "Sample statistics" and "Also termed as descriptive statistics". At the bottom of the slide, there is a footer with the IIT Guwahati logo on the left, the text "Indian Institute of Technology Guwahati" in the middle, and "Probability Distribution and Basic Descriptive Statistics" on the right.

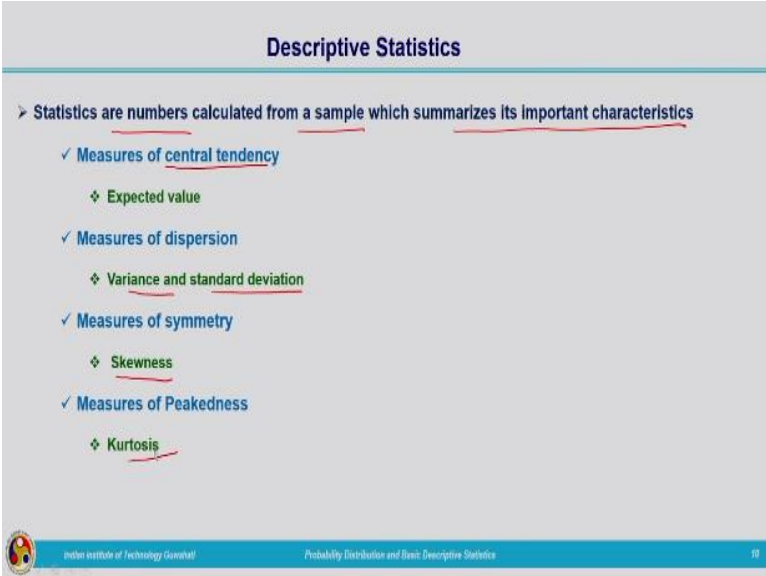
We are going to discuss about descriptive statistics. Why do we want to make use of these descriptive statistics in this particular topic of hydrology? Hydrologic analysis is involved with so much of data, that also data which consists of so much of uncertainties. So, we need to incorporate or we need to carry out the analysis in probabilistic manner. If we want to carry out the probabilistic analysis, we need to have understanding about the probabilistic character of the data, that is the probabilistic character of the random variables can be completely described if we are having the form of the distribution function, that is the type or form of the distribution function should be known to us or the associated parameter should be known to us, either the distribution function or the associated parameter should be known to us.

So, in such cases that is in the absence of the above parametric distribution, description about the population can be assessed if we are having some sample statistics, that is termed as descriptive statistics. If the variable is having or the variable is involved with certain uncertainty, it cannot take a single value, it will be taking up a value within certain range that is represented by the random variable, it can be continuous or discrete, that depends on the character of the variable which we are dealing with that.

So, the analysis regarding or involving these random variables will be possible only if we know the values taken up by the random variable and the corresponding probabilities. That is we need to have the relationship with the values taken up by the random variable and the associated probability that is represented by the probability distribution function.

So, in order to carry out the hydrologic analysis involving random variables, we need to have the idea about probabilistic distribution function or the associated variables. In the absence of these probability distribution function and associated parameters we may have to go for calculation of some statistics related to the data, that is we need to have some idea about the sample statistics which are termed as descriptive statistics.

(Refer Slide Time: 28:57)



Descriptive Statistics

- Statistics are numbers calculated from a sample which summarizes its important characteristics
 - ✓ Measures of central tendency
 - ❖ Expected value
 - ✓ Measures of dispersion
 - ❖ Variance and standard deviation
 - ✓ Measures of symmetry
 - ❖ Skewness
 - ✓ Measures of Peakedness
 - ❖ Kurtosis

Indian Institute of Technology Guwahati Probability Distribution and Basic Descriptive Statistics 10

Let us look into that. Statistics are the numbers calculated from a sample which summarizes its important characteristics. We are having a sample of data which is drawn from the population and for getting idea about that particular data we can summarize in terms of certain statistics, which will be giving us idea about the characteristics of the data, important characteristics we can assess from the descriptive statistics. Some descriptive statistics which we are going to explain here includes measures of central tendency, measures of dispersion, measures of symmetry and measures of peakedness.

So, four characteristics we are going to discuss here, if you are getting idea about these measures, important characteristics related to the concerned data set can be assessed. So, measures of central tendency includes expected value or mean. Actually, when we talk about central tendencies, from school days onwards we have studied mean, mode, median. So, what is meant by mean, mode, median, all these things you know already, here I am discussing only about the mean that is the expected value.

Now, second one is related to measures of dispersion, in this we will be discussing variance and the corresponding standard deviation. And coming to measures of symmetry, it includes skewness, measures of peakedness includes kurtosis. So, these are the four characteristics or properties in terms of descriptive statistics we are going to discuss.

(Refer Slide Time: 30:55)

The slide is titled "Measure of Central Tendency" and contains the following text:

- Mean
- Expected value is the first moment about the origin of a random variable
 - ✓ It is a measure of the 'central tendency' of a distribution
 - ✓ It is the location parameter

At the bottom of the slide, there is a logo on the left and the text "Indian Institute of Technology Guwahati" and "Probability Distribution and Basic Descriptive Statistics" on the right, with the number "11" in the bottom right corner.

First let us start with measures of central tendency that is nothing but our mean. It is termed as expected value, this is nothing but the first moment about the origin of a random variable. Random variable can take up different values within certain range, so if you are finding out the mean of this random variable about the origin that is termed as the expected value or mean.

Means also different ways we will discuss, I am talking about the mean here as the expected value that is the first moment about the origin. It is the measure of the central tendency of a distribution, it is also termed as location parameter, depending on the value of mean we can

understand the properties related to the central tendency, that is why it is also termed as location parameter.

(Refer Slide Time: 31:54)

Measure of Central Tendency

- An expected value or the mean of a random variable 'X'
- ✓ Population mean for a continuous RV,
 - ✦ Product of x and the corresponding probability density $f(x)$, integrated over the feasible range of RV
 - ✓ $E(x) = \mu = \int_{-\infty}^{\infty} xf(x)dx$ ✓
- ✓ Population mean for a discrete RV,
 - ✓ $E(x) = \mu = \sum_{i=1}^n x_i p(x_i)$
- ✓ The sample estimate of the mean is the average \bar{x} of the sample data
 - $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$

Indian Institute of Technology Guwahati | Probability Distribution and Descriptive Statistics | 9

So, an expected value or the mean of a random variable X can be written as following:

Population mean for a continuous random variable is expressed as the product of x and the corresponding probability density $f(x)$, integrated over the feasible range of random variable.

$$E(x) = \mu = \int_{-\infty}^{\infty} xf(x)dx$$

What we are doing, we are finding out the product of x and the corresponding probability density for each value of random variable. Random variable can take any value within certain range, so corresponding to each value there will be an associated probability and that is represented by $f(x)$, so we are finding out the product of x and $f(x)$, that is integrated within the range of $-\infty$ to $+\infty$ or within the required range, that is what is termed as the expected value or mean. In the case of population, it is denoted by μ .

Now population mean for a discrete random variable, the above integral expression is representing the expected value or mean corresponding to continuous random variable. For the case with the discrete random variable we will be having summation instead of integral. So,

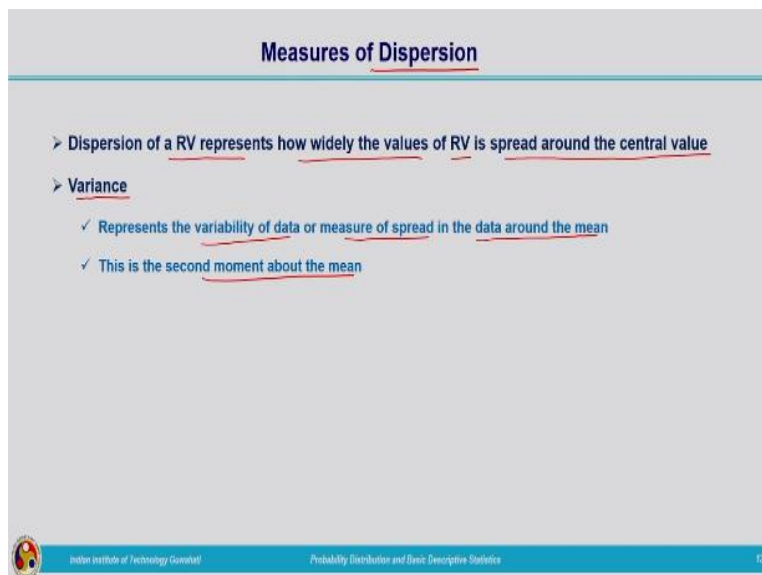
$$E(x) = \mu = \sum_{i=1}^n x_i p(x_i)$$

Now, coming to sample estimate, sample estimate of the mean is the average value (\bar{x}) of the sample data, that is represented by

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

This is applicable to the case with sample. In the case of population, we will be making use of these equations $E(x)$ for continuous random variable and for discrete random variable we will be using this equation. And that much about the central tendency represented by mean or expected value, that is representing the first moment about the origin.

(Refer Slide Time: 35:05)



The slide is titled "Measures of Dispersion" and contains the following text:

- Dispersion of a RV represents how widely the values of RV is spread around the central value
- Variance
 - ✓ Represents the variability of data or measure of spread in the data around the mean
 - ✓ This is the second moment about the mean

At the bottom of the slide, there is a logo on the left and the text "Indian Institute of Technology Gandhinagar" and "Probability Distribution and Basic Descriptive Statistics" in the center, with the number "33" on the right.

Now, we will move on to measures of dispersion, dispersion represents the variability in the data. So, dispersion of a random variable represents how widely the values of random variable is

spread around the central value. Central value is represented by the mean value and surrounding the mean value how the random variables are distributed or how widely these random variables are spread that idea we will be getting from the measures of dispersion, that is represented by means of term called variance.

Variance represents the variability of data or measure of spread in the data around the mean. We are having the mean value, around the mean how much of spread corresponding to the data points are there that can be understood by finding out the value corresponding to variance. This is second moment about the mean, you need to be careful, when we were discussing about expected value that was the first moment about the origin, here when we talk about variance and standard deviation, it is representing the second moment about the mean, it is not with respect to the origin, it is with respect to the mean value.

(Refer Slide Time: 36:34)

Measures of Dispersion

➤ **Variance**

- ✓ For a continuous RV

$$E(X-\mu)^2 = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$
- ✓ For a discrete RV

$$E(X-\mu)^2 = \sigma^2 = \sum_{i=1}^n (x_i - \mu)^2 p(x_i)$$
- ✓ The sample estimate of the variance is given by

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

❖ Here, the divisor is $(n-1)$ is used rather than using n to ensure that the sample statistics is unbiased

- Not having a tendency on average to be higher or lower than the true value

Indian Institute of Technology Gandhinagar Probability Distribution and Basic Descriptive Statistics 14

Now, let us see what are the expressions used for calculating this. Variance, for a continuous random variable can be expressed as

$$E(X-\mu)^2 = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Here, why square is coming? We are taking the second moment and also it is not with respect to the origin, it is with respect to the mean value, that is why we are taking the difference of $(x - \mu)$.

This is for the continuous random variable, for the discrete random variable we can write it as

$$E(X - \mu)^2 = \sigma^2 = \sum_{i=1}^n (x_i - \mu)^2 p(x_i)$$

Now, the sample estimate of the variance is given by

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

Now, here you can see in the denominator, there is an $(n - 1)$ term. This term is used in order to ensure that the sample statistics is unbiased, that is not having a tendency on average to be higher or lower than the true value, that is what is meant by unbiased statistics, that is it is not at a higher level or lower level compared to the mean, that is no bias is there.

Now, for expressing the measure of dispersion, we used to make use of another term termed as standard deviation in addition to variance. Variance is the second moment, that is it is having the dimension x^2 , what is the dimension of the random variable considered, square of that is the dimension of the case with variance. But in the case of standard deviation it is the under root of variance, so it will be having the dimension same as that of the variable.

(Refer Slide Time: 40:08)

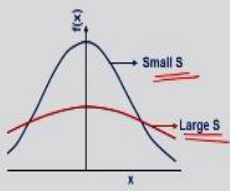
Measures of Variation

➤ **Standard Deviation**

- ✓ Measures the variability
- ✓ Square root of the variance

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

- ✓ Also known as shape parameter of a pdf
- ✓ The larger the standard deviation, the larger is the spread of the data



PDF with smaller and larger standard deviation with mean at zero

Indian Institute of Technology Guwahati Probability Distribution and Basic Descriptive Statistics 15

Standard deviation measures the variability. As in the case of variance this also represents the dispersion only but in this we are considering the square root of the variance. So, the standard deviation corresponding to the sample represented by S is

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

Similar way we can write the standard deviation for population also. This is also known as shape parameter of a probability density function. Now, from the standard deviation value, that is larger the standard deviation, larger is the spread of the data. This we can see with the help of a figure. We are going to plot the probability density function with smaller and larger standard deviation with mean at 0.

So, we are having the mean at 0 that is why the y-axis is going through the center, this is the probability density function and here you can see, in this case we are having small value of standard deviation, that is it is representing the smaller spread in the data, data is not spread in a vast way. But when you look at the second curve, you can see this is representing large value of standard deviation and also from the curve it is very clear that too much of spread is there within the data.

(Refer Slide Time: 41:50)

The slide is titled "Measures of Variation" with a handwritten note "Dispersion" next to it. The main heading is "Coefficient of Variability (CV)". Below this, there are two bullet points: "This is the dimensionless measure of variability" and "For population". The formula for population CV is given as $CV = \frac{\sigma}{\mu}$. Below this, there is another bullet point: "For sample". The formula for sample CV is given as $CV = \frac{S}{x}$. The slide footer contains the Indian Institute of Technology Guwahati logo and the text "Probability Distribution and Basic Descriptive Statistics".

Now, coming to another representation, that is in the case of variance it was having the dimension of x^2 and standard deviation we have taken the under root, so, it was having the same dimension as that of the variable which we are considering. For making it dimensionless we can consider the case with the coefficient of variability, CV. Coefficient of variability is a dimensionless measure of the variability or dispersion.

For population it is given by

$$CV = \frac{\sigma}{\mu}$$

For sample it is given by

$$CV = \frac{S}{x}$$

So, the dispersion or the variation in the data can be calculated by using coefficient of variability also. It is a dimensionless factor.

(Refer Slide Time: 43:08)

Measure of Symmetry

- Represents the symmetry of the distribution about the mean
- Quantifies the skewness in the data
- This is the third moment about the mean
- For continuous RV
$$E[(x-\mu)^3] = \int_{-\infty}^{\infty} (x-\mu)^3 f(x) dx$$
- For discrete RV
$$E[(x-\mu)^3] = \sum_{i=1}^n (x_i - \mu)^3 p(x_i)$$

Indian Institute of Technology Guwahati | Probability Distribution and Basic Descriptive Statistics | 17

Now, third measure that is measure of symmetric, that represents the symmetry of the distribution about the mean. We are having the distribution, we are having the mean value how this distribution or how the data is spread with respect to the mean, it is not representing the spread actually, it is representing the symmetry of the distribution with respect to the mean, it will be clear to you when we express it by means of the figure.

It quantifies the skewness in the data, how much skewness is there, with respect to mean whether it is skewed towards left or towards right when we plot the probability density function we can understand. So, this is nothing but the third moment about the mean. Expected value or mean was the first moment about the origin and standard deviation or the variance was the second moment about the mean and now we are talking about the skewness, it is the third moment about the mean.

For a continuous random variable this is the expression

$$E[(x-\mu)^3] = \int_{-\infty}^{\infty} (x-\mu)^3 f(x) dx$$

In the case of discrete random variable instead of integral we will be making use of summation. So, that can be written as

$$E[(x-\mu)^3] = \sum_{i=1}^n (x_i - \mu)^3 p(x_i)$$

In the case of discrete random variable we will be using the probability mass function.

(Refer Slide Time: 45:46)

Measure of Symmetry

➤ **Coefficient of Skewness**

- ✓ Skewness is made dimensionless by dividing with σ^3

$$\gamma = \frac{1}{\sigma^3} E[(x-\mu)^3]$$

- ✓ The sample estimate of coefficient of Skewness

$$C_s = \frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)S^3}$$

Indian Institute of Technology Gandhinagar Probability Distribution and Basic Descriptive Statistics 11

Now, coming to the coefficient of skewness. Skewness is made dimensionless by dividing with σ^3 . Skewness we have calculated in the previous slide that is the third moment with respect to mean, that is having a dimension of x^3 , that is third moment we have calculated. That has been made dimensionless by dividing it by σ^3 , σ is the standard deviation corresponding to population.

So, when we talk about coefficient of skewness γ , it is given by

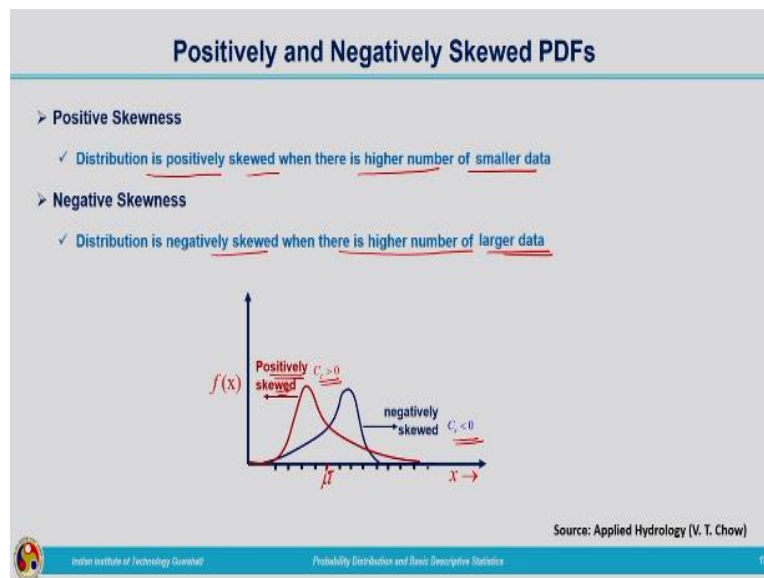
$$\gamma = \frac{1}{\sigma^3} E[(x-\mu)^3]$$

Now, coming to the sample estimate of coefficient of skewness, instead of σ we will be replacing it by s . So,

$$C_s = \frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)S^3}$$

So, we have seen the skewness, coefficient of skewness corresponding to population and also corresponding to sample.

(Refer Slide Time: 47:06)



Now, for understanding what is meant by skewness, how it is visually expressed, positive skewness and negative skewness. Positive skewness is the case where distribution is positively skewed when there is higher number of smaller data. In the data set we will be having different types of data, number of data corresponding to each will be different, in the data set we are having higher number of smaller data and the distribution will be skewed in that case.

So, it will be like this, we are plotting x along the x-axis and probability density function along the y-axis, in this you can see μ is marked here, μ is the mean or average value, expected value that will be showing the central tendency. So, when you compare this distribution with respect to μ this is skewed towards the right-hand side that is termed as positively skewed and $C_s > 0$. Positively skewed means we are having higher number of data corresponding to smaller values.

Now, coming to the other case that is negatively skewed. Negative skewness is representing the distribution is negatively skewed when there is higher number of larger data. When we are having higher number of larger data, if you are plotting the probability density function it will be looking like this, that is this is representing negatively skewed, that is $C_s < 0$.

In the case of positively skewed data set $C_s > 0$, it will be skewed towards the right. In the case of negatively skewed data, it will be skewed towards the left with coefficient of skewness $C_s < 0$. So, I hope now you got an idea what is meant by skewness, how the positively skewed data and negatively skewed data will be seen when we plot the distribution that is clear from this graph.

(Refer Slide Time: 49:31)

Measures of Peakedness

➤ Kurtosis

✓ Standardized fourth population moment about the mean

$$\beta_2 = \frac{E(X-\mu)^4}{(E(X-\mu)^2)^2} = \frac{\mu_4}{\sigma^4}$$

✓ where

- ❖ E - expectation operator,
- ❖ μ - mean
- ❖ μ_4 - the fourth moment about the mean
- ❖ σ - is the standard deviation

Indian Institute of Technology Guwahati | Probability Distribution and Basic Descriptive Statistics | 20

Now, next measure is measures of peakedness. Measure of peakedness is marked by means of kurtosis, this is a value which is representing the standardized fourth population moment about the mean. This is the fourth moment about the mean, expected value or the mean was the first moment about origin, variance is representing the second moment about the mean, skewness was representing the third moment about the mean and kurtosis is representing the fourth moment about the mean.

So, this is expressed by

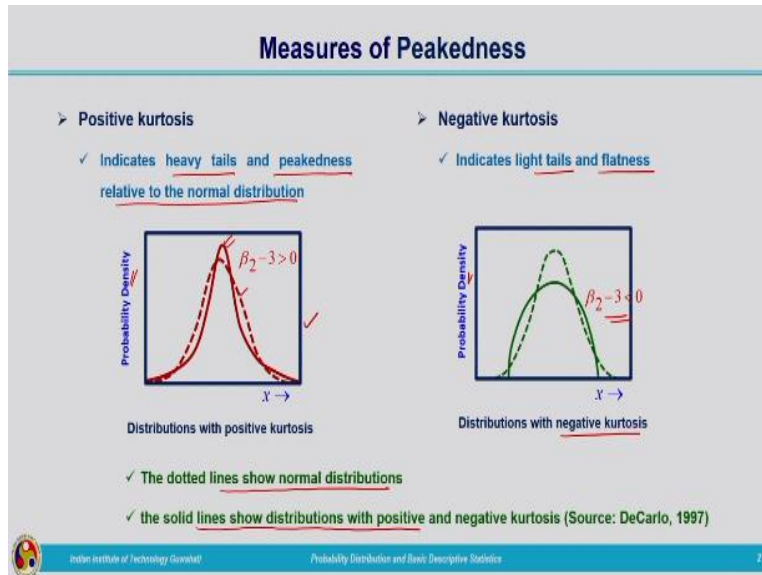
$$\beta_2 = \frac{E(X - \mu)^4}{(E(X - \mu)^2)^2} = \frac{\mu_4}{\sigma^4}$$

In this you know already

- ❖ E - expectation operator,

- ❖ μ - mean
- ❖ μ_4 - the fourth moment about the mean
- ❖ σ - is the standard deviation

(Refer Slide Time: 51:11)



Now, in this case also as in the case of skewness we are having positive kurtosis and negative kurtosis. Positive kurtosis indicates heavy tails and peakedness relative to the normal distribution. What is meant by normal distribution, we have not discussed about that, I will come to it when we start with the probability distribution and the negative kurtosis indicates the light tails and flatness. From the distribution plot we can understand whether high peakedness or flat peakedness is there, so that is what is represented by means of coefficient of kurtosis.

So, that can be plotted, distributions with positive kurtosis can be plotted like this. We are plotting the probability density along the y-axis and the specified values corresponding to random variables along the x-axis and this red line is representing the distribution corresponding to the given data. Second curve is marked with a dotted line and the dotted lines show the normal distribution and the solid lines show the distributions with positive kurtosis in this case.

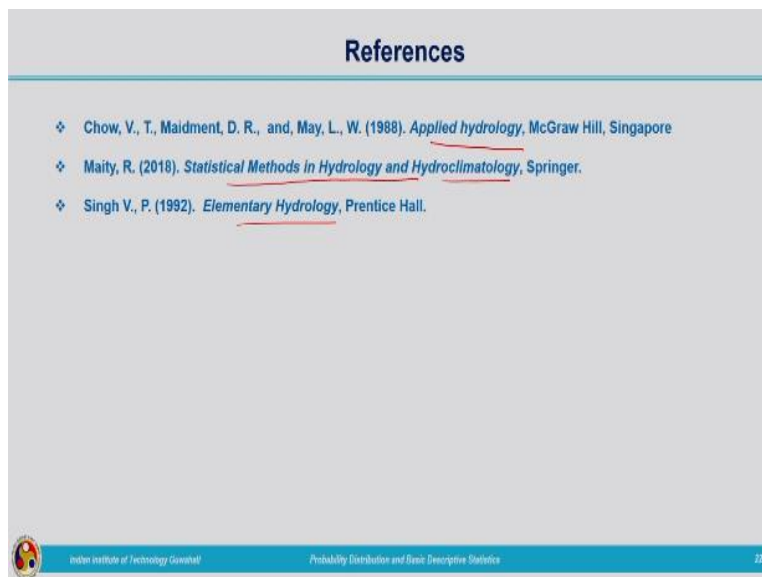
So, this solid line is representing the data with positive kurtosis, you can see the peak is high, peakedness is more compared to the normal distribution, this dotted line is representing the normal distribution, so for normal distribution we usually consider a kurtosis value of 3. So, for showing the peakedness it is represented by $\beta_2 - 3 > 0$, that is representing the positive peakedness or positive kurtosis value.

Now, coming to the negative kurtosis value, when we plot the distribution with negative kurtosis along the x-axis the specified values for the random variable and along the y-axis we are having the probability density and when we plot from this curve itself it is clear to you that the peak will be flat, it will be lower than that of the normal distribution.

So, that can be plotted like this. This dotted line is representing the normal distribution and the solid line is representing the distribution having negative kurtosis, in that case $\beta_2 - 3 < 0$. 3 is the kurtosis value corresponding to normal distribution, so, with that we are comparing.

So, that much about different descriptive statistics which we need to make use in further lectures.

(Refer Slide Time: 54:42)



For understanding these topics, you please go through the corresponding reference textbooks. These are some of the reference textbooks, different textbooks are Applied hydrology by Ven Te Chow and Elementary hydrology by Professor V. P. Singh. And in Professor Maity's textbook of

Statistical Methods in Hydrology and Hydroclimatology, you can get the details from the beginners level to the advanced level. Please refer through these textbooks and try to solve some of the examples related to the concepts which we have covered over here. Here I am winding up this lecture, thank you.