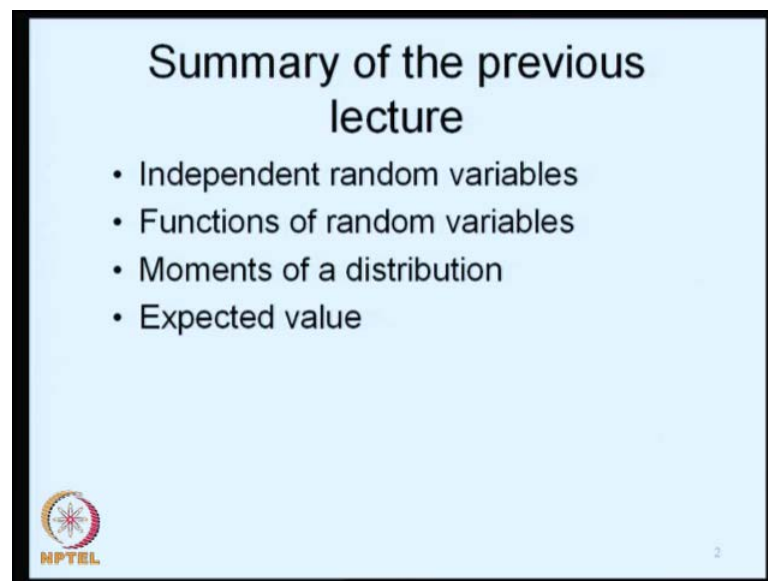


Stochastic Hydrology
Prof. P. P. Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

Lecture No. # 04
Moments of a Distribution

(Refer Slide Time: 00:27)



Good morning and welcome to this the fourth lecture of the course stochastic hydrology. If you recall what we have done in the last class, that is the lecture number three is that we introduced the concept of independent random variables, recall that we said two random variables are independent. If and only if, the joint density function f of x, y is given by the product of the marginal density functions g of x into h of y . Then we considered functions of random variables for initially we introduced functions of a single random variable, let say y is equal to h of x is a function of single random variable x , then we introduced how to compute the how to estimate the pdf of y given the pdf of x .

Then we went on to consider functions of two random variables given f of x, y there is a joint density function of x and y , how to compute the joint density of u comma v , where u and v are functions of x and y . They are continuous function, and we have seen the examples for continuous functions. Then towards the end of this lecture we introduce the

moments of a distribution, and introduced a concept of expected value. The last two points will just revise again today moment of distribution and expected value, and then move on.

(Refer Slide Time: 02:02)

Moments of a distribution

n^{th} moment about the origin

$$\mu_n^o = \int_{-\infty}^{\infty} x^n f(x) dx$$

$E(X)$: Expected value of 'X'
: First moment about the origin

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

n^{th} moment about the expected value

$$\mu_n = \int_{-\infty}^{\infty} (x - \mu)^n f(x) dx$$

The slide also features a graph of a probability density function $f(x)$ versus x . A vertical line marks the expected value $E(x)$, and a small shaded area under the curve is labeled dx . In the bottom right corner, there is a small video inset of a man in a white shirt speaking.

So, if you recall we said the n moment about the origin this is the origin and we are talking about the moments of pdf, f of x this is x the n the moment is given by integral minus infinity to plus infinity x to the power n f of x dx . So, we are essentially taking the moment of this area about this point. The first moment that is when n is the is equal to one the first moment about the origin this 0 here represents that we are taking the moments about the origin. The first moment is called as the expected value of x and by putting n is equal to one here we get expected value of x is equal to minus infinity to plus infinity x f of x dx .

So, here we have put n is equal to 1 this is also denoted by μ and mean expected value are one at the same. So, we call it as the mean of the distribution mean of the random variable we defined the expected value we have fixed the expected value like this here. We start taking moments about the expected value itself and we define the moment about the expected value as μ_n we do not put the o there or 0 there. So, if we only write μ_n it implies that we are taking the moments about the expected value and that is defined as minus infinity to plus infinity x minus μ to the power n f of x dx where f of x is the pdf

of x . This will define the n about the expected value then using this moments we define several measures which will give the properties of random variable.

(Refer Slide Time: 04:06)

Measures of central tendency

Mean:

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

Discrete case: $\mu = \sum_{i=1}^n x_i p(x_i)$ n : Sample size

Sample estimate of the mean: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Mode: Value with highest frequency of occurrence

Median: Value such that 50% of area is on either side

NPTEL

For example, how the values are distributed around the mean what kind of central value we can expect in terms of the mean or average value and the maximum, the value with higher frequency and the value that divides the distribution into two parts and so on. So, first we define measures of center tendency and the most important measure of center tendency is the mean, which we just define. So, μ is given by minus infinity to plus infinity $x f x dx$ in the case of continuous case continuous random variables. This is a measure defined on the population then in the case of discrete random variables or when we have random variable taking on only discrete values with associated probability mass function.

We define the μ as summation i is equal to 1 x_i which is of i the discrete random variable into p of x_i which is the probability with which the value x_i occurs and n here is the sample size or n is the number of values or n is the number of values that the random variables discrete random variables x can assume. Now, the sample estimate of the mean remember as I told in the last class from the population we move on to samples, which are the actual values that the random variable as taken or sample is actually a subset of the population, and if you have n numbers of values in the sample.

The sample estimate of the mean is given by simply the arithmetic average \bar{x} is equal to $\frac{1}{n} \sum_{i=1}^n x_i$ and this summation is from i is equal to 1 to n where n is the sample size.

So, this is the most important measure of center tendency mainly the mean, we also have the concepts of mode and the median mode is that value which occurs with the highest frequency.

(Refer Slide Time: 06:40)

Moments of a distribution

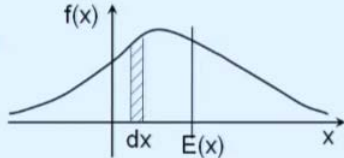


n^{th} moment about the origin

$$\mu_n^o = \int_{-\infty}^{\infty} x^n f(x) dx$$

$E(X)$: Expected value of 'X'
: First moment about the origin

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

n^{th} moment about the expected value

$$\mu_n = \int_{-\infty}^{\infty} (x - \mu)^n f(x) dx$$




For example, if you have a distribution like this, that you have let say a distribution something like this and you have the highest frequency value here this defines the mode.

(Refer Slide Time: 07:06)

Measures of central tendency

Mean:



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

Discrete case: $\mu = \sum_{i=1}^n x_i p(x_i)$ n : Sample size

Sample estimate of the mean: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

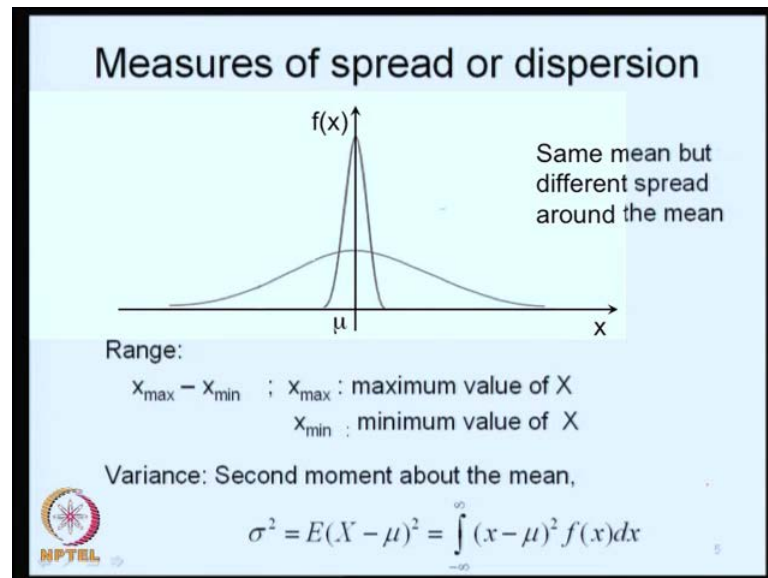
Mode: Value with highest frequency of occurrence

Median: Value such that 50% of area is on either side

And median is that value, which divides the distribution into two halves; that means, the area towards the right of the median is equal to area towards the left of the median, both of which are equal to 50 percent of the area or most. In most hydrologic analysis we use the concept of the mean more often than the concept of the mode and the median it is important for us to understand the concept of mean.

(Refer Slide Time: 07:35)



Then we have the measures of spread or dispersion let us look at a sample consisting of values 0, 10, and 20 let us say there are three values in the sample. What is the sample estimate of the mean? It would be simply the arithmetic average and therefore, it would be 10 we have another sample which has again three values, but the values are 9, 10 and 11 again the sample estimate of the mean is 10, which is the arithmetic average. Now both of these have the same mean, but as you can see the spread around the mean is much higher in the first case compare to the spread around the mean compare to the of the second case and therefore, it is important for us to see how the values are spread around the mean or how they are dispersed around the mean.

And therefore, we introduce a measure for measuring the spread of the distribution around the mean the most obvious measure is the range, which is given by simply the difference between the maximum value and the minimum value. So, it indicates actually how far are the values spread not necessary around the mean in this case $x_{\max} - x_{\min}$ will give you the range of values that you may expect, but the more important

measure of spread is the variance, which is based on the second moment about the mean which is in fact, equal to the second moment of the about the mean.

So, sigma square which is the variance is defined as expected value of x minus mu the whole square and recall that the expected value of a function is given by that function multiplied by the integral of that function multiplied by f of x dx. So, in this case we are looking at the expected value of x minus mu the whole square and therefore, this is given by integral minus infinity to plus infinity x minus mu the whole square f of x dx. So, as you can expect here for example, the spread of this distribution is much higher compare to the spread of this distribution. So, the variance that you can get you get from this distribution is higher than will be higher than the variance from this distribution. So, the variance actually indicates how the values are spread around the mean. The higher the variance the higher is the spread.

(Refer Slide Time: 10:36)

Measures of spread or dispersion

Sample estimate - Variance :

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

n: No. of observations in the sample

Standard deviation:

$$\sigma = +\sqrt{\sigma^2}$$

Positive squareroot

$$s = +\sqrt{s^2}$$



Coefficient of variation:

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

Population

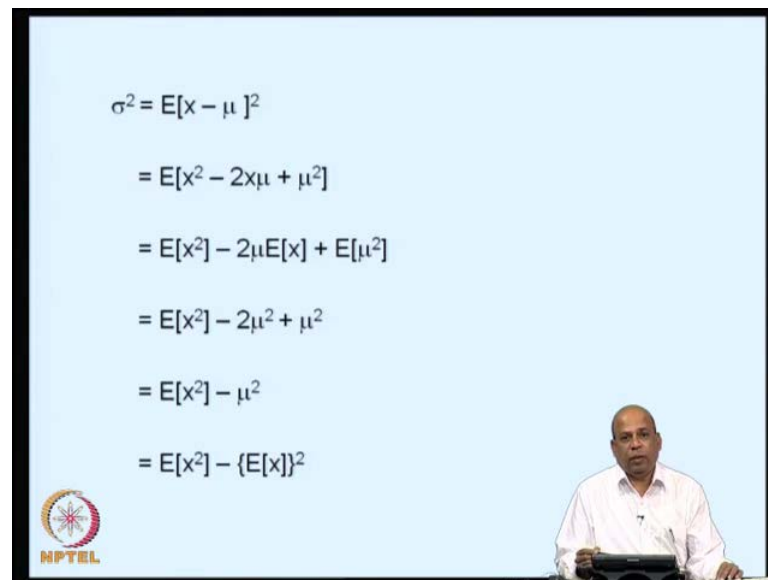
$$= \frac{s}{\bar{x}}$$

sample space

The sample estimate for the variance again recall that we have defined this, for the population this definition; is for the population, and the sample estimate for that is given by summation i is equal to 1 to n x i minus x bar to the whole square divided by n minus 1.

(Refer Slide Time: 11:02)


$$\begin{aligned}\sigma^2 &= E[x - \mu]^2 \\ &= E[x^2 - 2x\mu + \mu^2] \\ &= E[x^2] - 2\mu E[x] + E[\mu^2] \\ &= E[x^2] - 2\mu^2 + \mu^2 \\ &= E[x^2] - \mu^2 \\ &= E[x^2] - (E[x])^2\end{aligned}$$

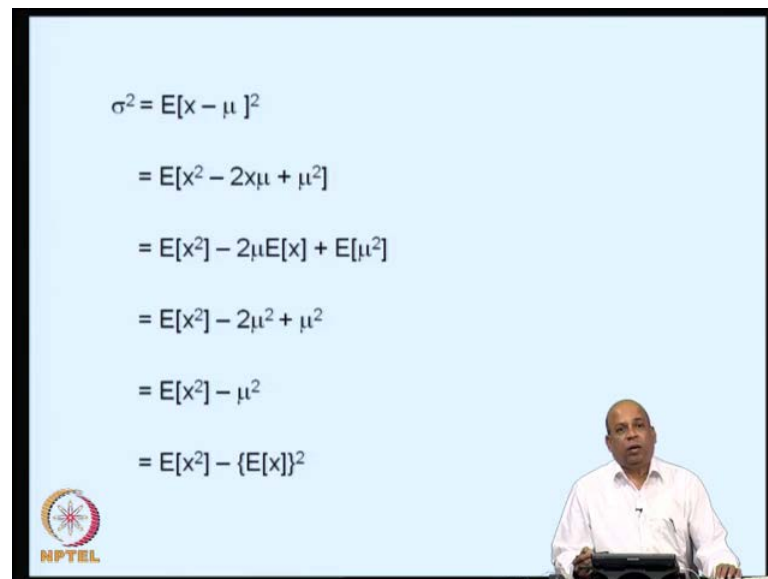
In the case of the mean the sample estimate was given by you look at this point \bar{x} summation of x y over n . So, n is the sample size. why is it that we use n minus 1 here of course, you can also use n as one of the estimates, but we use n minus one for getting what is called as an unbiased estimate of the variance will come to the parameter estimation subsequently in this course, but right now you remember it is not n , but n minus 1 because we have looking at an unbiased estimate of the variance. Then we define the standard deviation as the positive square root of the variance. So, the standard deviation is simply the positive square root of variance and the associated sample estimate for that is given by s as the positive square root of s square.

Now if you have a sample let say of stream flows, stream flows measured in million cubic meters then what are units of \bar{x} or the arithmetic average which is a sample estimate of the mean. The \bar{x} will also have the same units as the variable x itself. So, in this case \bar{x} will have the units of million cubic meters. Similarly the standard deviation you look at this point x_i minus \bar{x} whole square. So, \bar{x} has the units of the flows in this example x_i has units of flows. So, x_i minus \bar{x} the whole square has the units of flow square or million cubic meter square and when we take the square root the standard deviation will have same units as the original variable.

So, the standard deviation in this case will have variables of million cubic meter, but when we bound to compare two distribution different two distributions of different

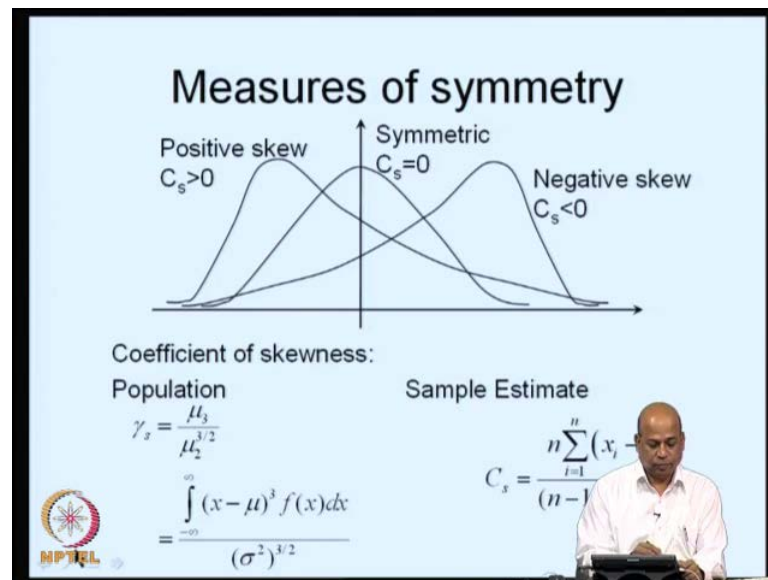
random variable, let say we want to compare the distribution of rainfall as well as the stream flow. We want to have a measure of the spread which is independent of the units and therefore, we define what is call as coefficient of variation, and we define that as sigma over mu as you can see sigma has same units as x mu has the same units as x and therefore, c v will not have any units, and for the sample it is given by s by \bar{x} . So, in general we use coefficient of variation as a measure of spread when we want to have comparison with comparison among different distributions.

(Refer Slide Time: 14:00)


$$\begin{aligned}\sigma^2 &= E[x - \mu]^2 \\ &= E[x^2 - 2x\mu + \mu^2] \\ &= E[x^2] - 2\mu E[x] + E[\mu^2] \\ &= E[x^2] - 2\mu^2 + \mu^2 \\ &= E[x^2] - \mu^2 \\ &= E[x^2] - (E[x])^2\end{aligned}$$

So, we have defined sigma square as expected value of x minus μ the whole square all these have to be capital here say for example this x , because we are talking about a random variable.

.(Refer Slide Time:14:18)



This is a capital random variable, so this X , similarly this is X so, from here expected value of x minus μ the whole square we expand this. So, we are talking about expected value of x square minus two $x \mu$ plus μ square. So, when we simplify this recall the properties of the expected value. So, I can write this as expected value of x square minus two μ being constant it comes out into expected value of x plus expected value of μ square this will be written as expected value of μ square minus two μ into expected value of x again μ .

So, this will be minus two μ square plus μ square. So, this will be equal to expected value of μ square minus μ square which is written as expected value of x square minus expected value of x the whole square, this is a useful result which in many situations we use in many application we use this particular expression. So, we defined two measures now measures of central tendency which will indicate how the distribution is in terms of the central values example the mean the mode the median etcetera, then we define a measure of the spread or dispersion which use how far the value are spread for the mean the variance is a important measure of dispersion.

Now, we look at another important measure of the distribution which gives us whether the values are symmetrically distributed or there is a skewness there, and if there is a skewness whether it is a positive skew or negative skew. So, we introduce a measure of symmetry which is again depend which is depend on the moments from the first moment

about the origin which provided you the expected value of x . We went on to the second moment about the expected value about the mean and define the variance which is expected value of x minus μ the whole square which is in fact, the second moment about the like this we keep proceeding to higher order moments we consider the third moment about the expected value, and then define a measure go to the higher order fourth moment and then define a associated measure like this we are keep in moving to higher order moments about the expected value each moment has some implications on the type of the distribution.

So, as you consider higher and higher moments you get better and better idea about the distribution itself. So, let us look at the measure of symmetry now what is that we are looking at we are looking at let say there is a distribution something like this there is a long tail towards the right an another distribution, which have the long tail towards the left and has a similar peak as any of them and there is, another distribution which is perfectly symmetrical about certain point. We define the coefficient of skewness based on the third moment now whenever we are defining the measures we normalize those measures the associated moment with the standard deviation. What is a standard deviation? Standard deviation is μ to the power half, because μ^2 is a variance and to the power half which is a square root to the standard deviation.

So, when you take the third moment you standardize you normalize that with the standard deviation to the power three when you take the fourth moment you normalize that with the standard deviation to the power four and so on. So, the coefficient of skewness we define here with the third moment μ_3 divided by μ to the power $3/2$ remember μ to the power $3/2$ is the standard deviation. So, μ to the power $3/2$ is the standard deviation cube.

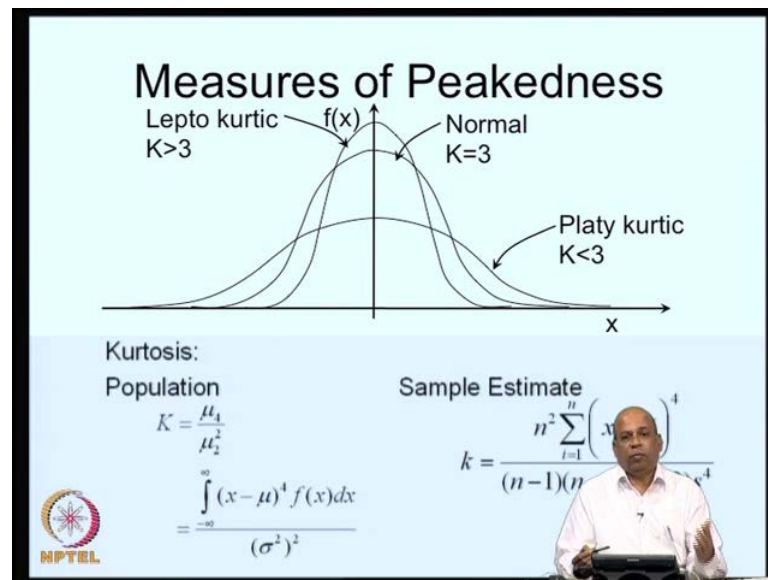
So, this is written as this is, your μ_3 $\int (x - \mu)^3 f(x) dx$ recall that μ_n is $\int (x - \mu)^n f(x) dx$ therefore, this term here indicates the third moment divided by σ^3 which is μ^2 to the power $3/2$ which means σ to the power 3 from this we can write the expressions for the sample estimates. So, for the sample estimates we replace the integral with the summation is equal m to n x_i minus \bar{x} we written the same power as the moment x minus μ the whole cube. So, we have written the same moment. Then, for the standard deviation what is it that we had n to the power 0 and x_i minus \bar{x} the whole square. We had there for the variance now

we start increasing the power in the numerator power of n in the numerator and start increasing the terms in the denominator for the variance we had n to the power 0 and here we had n minus 1.

For the skewness we put n to the power one and increase another term here n minus 1 n minus 2 and this s cube correspond to the sigma cube here. So, we are essentially normalizing this term with the standard deviation same power. So, for example, if you go to the higher moment, next higher moment you will be writing for μ_4 that is μ_4 and therefore, the denominator will be $\mu_4 \mu_2$ to the power 4 by 2 which is μ_2 square and therefore, what we will write here we will write for the fourth moment we will write n square and then n minus 1, n minus 2, n minus 3 and this power becomes 4. Like this from one moment to next higher moment we keep increasing the order of the moment and normalize that with respect to the standard deviation raise to the same power.

The coefficient of symmetry or coefficient of skewness as we have defined here will indicate whether the values are symmetrically distributed or there is a negative skew in which case the c_s has defined here will be negative or that has a positive skew in which case the c_s has defined here will have a positive value. Note that for the positive skew the values are stretched in the sense the distribution is stretched to the right. So, it is skewed to the right and for the negative skew there is a long tail to the left. So, that if you have a long tail to the right it is called as the positive skew you have a long tail to the left it is called as the negative skew.

(Refer Slide Time: 22:39)



Then we go to the another measure called as measure of peakedness. So, what is it that we did first we define how centrally the values are distributed then how far they are spread from the mean, and we also saw how symmetric or how skewed is the distribution of the values. So, these 3 give us some idea about the distribution itself, but there is another important measure which is defined with respect to a called normal distribution. So, if you have a normal distribution you may have something like this.

So, how peaked is a distribution that means, you look at this distribution for example, this is much flatter compare to the normal distribution, and then there may be another distribution which has the same spread, same mean, same mode, same median, etc, and same skewness coefficient also it may have, but it have a different peakedness. So, the values are the distribution have a much higher peak compare to the So, called normal distribution to measure this that means, how peaked is the distribution we introduce what is called as a kurtosis it also called as a kurtosis coefficient or coefficient of kurtosis, but in this course we will simply denote this as kurtosis.

So, as said we move on to higher moment from the previous one in the previous one, we consider the third moment about the mean we consider now the fourth moment about the mean. So, the kurtosis is defined based on the fourth moment of auto mean and then we standardize this with respect to $2 \mu^2$ to the power 4 by 2. μ^2 to the power half is the standard deviation and that we are raising it to the power 4 μ to the power 4 by 2 will

give you μ to the power 2. So, kurtosis is given by μ^4 divided by μ^2 to the power 2, which if we write this is again population this would be minus infinity to plus infinity and μ^4 here which is $\int (x - \mu)^4 f(x) dx$ divided by μ^2 to the power 2 which is μ^2 is σ^2 which is a variance to the power 2.

The associated sample estimate is written as recall that in a previous measure which has measure of symmetry we had used μ^3 there we had n here. So, we go on to the higher power of n and write n^4 and this power is the same as the power of the moment. So, $\sum (x_i - \bar{x})^4$ divided by n^4 here we increase one term $n - 1$, $n - 2$, $n - 3$, and here we have got σ^4 and therefore, we write it as s^4 . So, this is a sample estimate of a kurtosis. So, together these measures mainly the measure of center tendency measures of dispersion. In fact, and the measure of skewness or the measure of symmetry which is given by the coefficient of skewness and the measure of peakedness together all of this will give us an idea of how the distribution is likely to be. So, if you have a sample let say that over the last 50 years you have observed the stream flow at the particular location.

So, from this sample you should be able to estimate you should be able to determine the sample estimates of the mean or the expected value the sample estimate of the variance and therefore, the standard deviation from which you can get the coefficient of variation and also estimate the coefficient of skewness or the coefficient of the symmetry and the coefficient of kurtosis are in this particular case, we call it as kurtosis simply call it as kurtosis.

(Refer Slide Time: 27:16)


Example-1

Consider the pdf

$$f(x) = \begin{cases} 3x^2 & 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Obtain

1. $E(X)$
2. $E(3X-2)$
3. $E(X^2)$



10

So, these measures will give us other side or an idea about which type of distribution this sample may fit let us take an example. Numerical example first we will work with expected value along, let say you have a pdf, f of x is equal to $3x^2$ and which is defined for x varying between 0 and 1, let us obtain the expected value of x and the expected value of a function defined on x which is three x minus 2 and the expected value of x^2 , which is another function defined on x .

(Refer Slide Time: 27:50)

Example-1(contd.)

1. $E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 x \cdot 3x^2 dx = 3 \left[\frac{x^4}{4} \right]_0^1 = \frac{3}{4}$
2. $E(3X-2) = \int_{-\infty}^{\infty} (3x-2)f(x)dx = \int_0^1 (3x-2) \cdot 3x^2 dx$
 $= \int_0^1 (9x^3 - 6x^2) dx = \left[\frac{9x^4}{4} - 2x^3 \right]_0^1 = \frac{1}{4}$
3. $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 x^2 \cdot 3x^2 dx = 3 \left[\frac{x^5}{5} \right]_0^1$



So, the expected value of X is simply given by minus infinity to plus infinity $x f$ of $x dx$ and, because f of x is defined to be 0 between the range 0 and 1, we integrate between 0 and 1 this is x this is f of x $3 x$ square with respect to x . So, we get it as 3 by 4 . So, the expected value of x is 3 by 4 this is also mean and it is also denoted by μ then the expected value of any function of x is given by integral minus infinity to plus infinity that particular function into f of $x dx$.

So, expected value of $3 x$ minus 2 is given by minus infinity to plus infinity $3 x$ minus 2 into f of x which is $3 x$ square into dx and when you integrate that you will get one by 4 , similarly the expected value of x square is given by minus infinity to plus infinity x square f of $x dx$ this is the function x square is the function and this is pdf and this you get it as 3 by 5 . Recall that expected value of x square can be used to get the variance once you have expected value of x we wrote expected value that is sigma square for example, sigma square we wrote is as expected value of x square minus expected value of x the whole square. So, we can get the variance from expected value of x and expected value of x square.


(Refer Slide Time: 29:40)

Example-2

Obtain the sample estimates of mean, standard deviation, coefficient of variation, coefficient of skewness and kurtosis for the following observed data of annual stream flow for 15 years.

Year	1	2	3	4	5	6	7	8	9	10
Avg. yearly stream flow (Mm ³)	150	129	160	152	165	138	149	115	97	154

Year	11	12	13	14	15
Avg. yearly stream flow (Mm ³)	168	110	108	105	125



We will take another example where we have a sample the previous example consisted of population estimate for example, this is a expected value based on the pdf and for the population, now we will go to an example where we have a sample of values for example, average yearly stream at a particular location. . So, this have been measured for

last 15 years although I must alert, you that using the sample estimates then using the sample estimates for the population as approximation of population measure we must have larger number of values here larger number of years of observation not just the 15 years, but as an example we take 15 years long. So, these are the observed value 150 129 and So, on. So, these are given for the last 15 years. So, when you have observed values and you want to have an idea of distribution you use the sample estimates for the measures that we just defined mainly the \bar{x} which is a estimate for the men x square which is an estimate for the variance sigma square and the coefficient of skewness which is denoted as c_s for the sample estimates and the kurtosis which is k .

(Refer Slide Time: 31:07)



Example-2 (contd.)

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

$$\sum_{i=1}^n x_i = 150+129+160+152+165+138+149+115+97+154+168+110+108+105+125 = 2025$$

Therefore mean, $\bar{x} = 2025/15 = 135 \text{ Mm}^3$

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

So, first we get the sample estimate for the mean as simply the arithmetic average \bar{x} by n simply add up all the values divided by the number of values which is 15 in this case. So, you get the sample mean as 135 million cubic meter and the variance is given by $x_i - \bar{x}$ the whole square $n - 1$. So, that once you get the sample mean or the \bar{x} you open out a table tabular column, because you need all these powers $x_i - \bar{x}$ the whole square $x_i - \bar{x}$ the whole cube and so on you need all these columns so we open out tabular column like this.

(Refer Slide Time: 31:49)

Year	Avg. Stream flow Mm ³ (x _i)	(x _i - \bar{x})	(x _i - \bar{x}) ²	(x _i - \bar{x}) ³	(x _i - \bar{x}) ⁴
1	150	15	225	3375	50625
2	129	-6	36	-216	1296
3	160	25	625	15625	390625
4	152	17	289	4913	83521
5	165	30	900	27000	810000
6	138	3	9	27	81
7	149	14	196	2744	38416
8	115	-20	400	-8000	160000
9	97	-38	1444	-54872	2085136
10	154	19	361	6859	130321
11	168	33	1089	35937	1185921
12	110	-25	625	-15625	390625
13	108	-27	729	-19683	531441
14	105	-30	900	-27000	810000
15	125	-10	100	-1000	10000
Σ	2025	0	7928	-29916	

So, this column gives you the year number and then these are the values. So, actual x i values. So, this is xi equal to 1, xi is equal to 2 etc, and these are the x i values x 1 is equal to 150 x 2 is equal to 129 and so on. So, sum it over you get 20 25 and 20 25 by 15 gives you the mean which was 135 as obtained before.

(Refer Slide Time: 32:19)

Example-2 (contd.)

Variance, $s^2 = \frac{7928}{15-1} = 566$

Standard deviation, $S = +\sqrt{s^2} = 23.8 \text{ Mm}^3$

Coefficient of variation, $C_v = S/\bar{x} = 23.8/135 = 0.176$

Coefficient of skewness, $C_s = \frac{n \sum (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$

$$= \frac{15 \times (-29916)}{(15-1)(15-2)23.8^3}$$

$= -0.183 < 0$, negatively skewed

So, this is your arithmetic mean so, we use this 135 and then get x i minus x bar which is 150 minus 135 which is 15 remember that because x bar can be greater than x i, you can have negative values here that would be some negative values. In fact, a first order check

of whether you are doing your computations is that the summation of the first order deviation which is $x_i - \bar{x}$ must be equal to 0, then we get $x_i - \bar{x}$ the whole square which is simply this term square minus \bar{x} square 25 square and so on.


Now, this you add it up you get the variance. So, how do we get variance this is summation of $x_i - \bar{x}$ the whole square divided by $n - 1$. So, this is summation of $x_i - \bar{x}$ the whole square divided by $n - 1$ which is 15 minus 114 that is how you get the variance, similarly, we go to coefficient of skewness you add up $x_i - \bar{x}$ the whole cube, because the order of the power is odd you will get negative value use here. So, there will also negative value not only the negative values and then you sum it over and then divided by you use your sample estimate formula you get the coefficient of skewness.

Then you go to the $x_i - \bar{x}$ whole cube which is whole to the power 4 which is these values raise to the power 4 sum them over you get fourth order the fourth order power. fourth power of the deviation $x_i - \bar{x}$ use these to get the sample estimates for example, sample estimates for the variance is $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$ this is i is equal to 1 to n , similarly then you take the standard deviation which is the positive square root of the variance you get 23.8 million cube meter then we get the coefficient of variation which is s by \bar{x} recall that we defined coefficient of variation to obtain a unit less measure of the spread or the deviation.

So, you get s by \bar{x} as 0.176 then coefficient of skewness we use the third power of deviation $x_i - \bar{x}$ the whole cube divided by $n - 1$ $n - 2$ s to the power 3 we have got s here and then that s we use here and this minus 16 is obtained from the summation of this term here $x_i - \bar{x}$ cube. So, this is minus 2916 we use that and then the standard deviation to the power 3. So, we get c_s as minus 1 minus 0.183 remember that the coefficient of skewness can be negative. So, when it is negative we get a negatively skew distribution.

(Refer Slide Time: 35:52)

Example-2 (contd.)

$$\begin{aligned}\text{Coefficient of Kurtosis, } k &= \frac{n^2 \sum_{i=1}^n (x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} \\ &= \frac{15^2 \times 6678008}{(15-1)(15-2)(15-3)23.8^4} \\ &= 2.14 \\ &< 3, \text{ Platy kurtic}\end{aligned}$$
16

Then we get also the coefficient of kurtosis which is also called as simple kurtosis k is equal to n square which is 15 square and $\sum (x_i - \bar{x})^4$ whole to the power minus 4. In this case it was 6678008 which is the summation of this column here and we use that raise as standard deviation to the power 4, you have $n-1$, $n-2$, $n-3$ which 14, 13 and 12 then you get 2.14 if the kurtosis is less than 3 then, it called as platy kurtic and it generally indicates flat distribution. For example, a platy kurtic distribution will have some shape something like this if you this was your normal distribution this would be k is equal to 3 recall that we define the normal distribution like that if k is than 3, you may expect a flat distribution like this. So, lower the k the flatter will be the distribution. So, in this particular case you get 2.14 as the kurtosis which is less than 3 and therefore, it is the platy kurtic distribution.

(Refer Slide Time: 37:12)



So what? It is that we did. So, far that if we have a sample of observe value we first get an idea of what kind of distribution we may expect this particular sample to follow this we did with several measures measure of center tendency measure of soft spread, how far the values spread from the mean, and then how symmetric is the distribution is there is a symmetry in the distribution or there is a lack of symmetry if there is a skewness whether it is positively skewed or negatively skewed then how peaked is the distribution with respect to some normal normally defined distribution and so on.

So, this gives from the sample of values that we have we get a idea of what kind of distribution we may expect for the sample to follow, now from several applications there is a large number of distributions, which have defined for hydrologic application, and also for other applications recall that when we introduce the concept of pdf probability density function we said any functions satisfies those condition; for example, effects must be non negative effects must be greater than equal to 0 and the integral minus infinity to plus infinity F of X t X must be 1. So, any function satisfies these two conditions is a potential condition for being a pdf in the continuous case that is a continuous random variable.

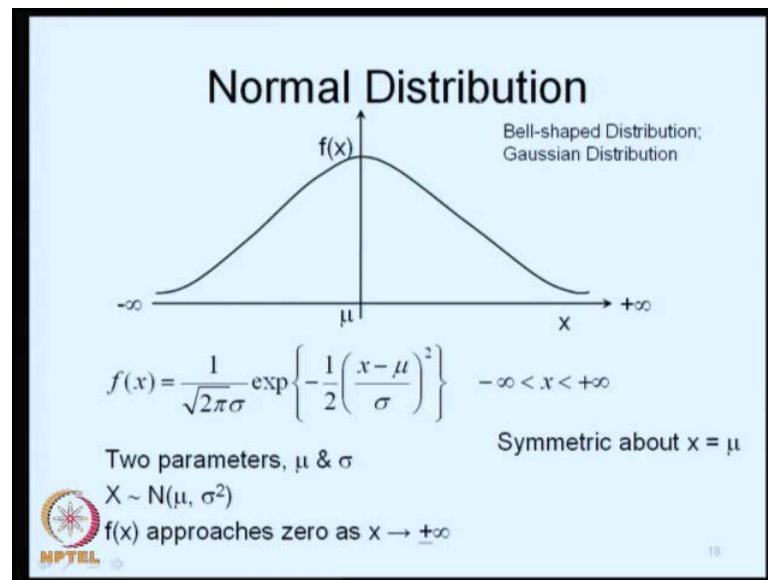
So, there are a handful of distribution which are commonly used in hydrology. We will see some of them the idea here is the if we want to use that, let say have a data for last 20 years of stream flow and you have estimated all these estimated measures for example, X

bar you have X^2 you have estimated coefficient of X skewness you have estimated coefficient of peakedness you have estimated from these you would like to use well defined distribution. So, we will go through some of the distributions that share commonly used in hydrology, and we will also see some applications where particular type of distribution is used for a particular type of process. For example, if we have a daily rainfall what kind of distributions normally used and monthly stream flow seasonal flow when the time scales are large processes are more or like normal processes what kind of distributions we use and so on.

So, we introduce some distributions today and may be continue to the next class also this distributions are commonly used for several hydrologic application the most commonly used distribution is the normal distribution in fact, the normal distribution in hydrology we use it for most. So, called normal processes for example, monthly stream flow seasonal rainfall the smoothen processes which are aggregates over large amount of time and the normal distribution is also called as the bell-shaped distribution and after the German mathematician who introduced this distribution it is called as the distribution in fact, the normal distribution was first introduced for analysis of errors from the observed data, and the similar data or experimental data.

So, we used the normal distribution as I said in most cases where we have aggregation of values over a sum more than process for example, if we are talking about monthly stream flows seasonal rain fall or seasonally re-precipitation where the values can be considered as summation of smaller processes large number of smaller processes.

(Refer Slide Time: 41:06)



We can use normal distribution as a fast and accurate approximation. The normal distribution is also the most commonly used distribution in many of the hydrologic applications where we are dealing with normal processes. Again, I repeat for example, we have the looking at the inflows to a reservoir, the storage levels on the reservoir operated based on a certain operating policy. So, how the reservoir levels are fluctuated, now such normalized processes which are aggregate of smaller processes happening over large over a fairly reasonable large amount of time we can approximate them using normal distribution.

Now the normal distribution has very interesting properties, because of which, it becomes the most commonly used distribution. First, it is a perfectly symmetrical distribution and it is symmetrical about the mean μ . The random variable is X , the pdf f of x for the normal distribution is given by $\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$ and this is defined for x varying between minus infinity to plus infinity as you can see from the distribution. This is the pdf, it has two parameters μ and σ . So, once you have defined parameters μ and σ , this pdf is completely defined.

So, it is called as it has two parameters μ and σ and we denote it as X followed by normal distribution with two parameters μ and σ . We use this notation. So, when we use this notation, it means that X is a random variable which follows the normal distribution with parameters μ and σ^2 . σ^2 is a variance. So, many

times you may see that the notation consist of mu n sigma, it does not matter it only indicates that instead of variance we are using the standards deviation the distribution approaches 0 asymptotically as x goes to minus infinity or plus infinity and it is symmetric about x is equal to mu the mean the mode and the median are all the same. So, x is equal to mu also defines the mean and the median, and the mode because this is also the highest frequency value.

(Refer Slide Time: 44:38)

Normal Distribution


Coefficient of skewness, $\gamma_s = 0$

Kurtosis, $K = 3$

$Y = a + bX$ – Linear function of 'X'

$Y \sim N(a+b\mu, b^2\sigma^2)$

$$F(x) = \int_{-\infty}^x f(x)dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \quad -\infty < x < +\infty$$

 19

Then there are other interesting properties of normal distribution, because it is perfectly symmetrical you have a gamma S which coefficient of skewness to be 0, recall that if it is symmetric distribution you get gamma S as 0. So, normal distribution has a coefficient of 0 and kurtosis is actually defined with respect to normal distribution is 3, but the useful result of normal distribution is that if x is normal distribution distributed with mu and sigma that is we are talking about x here as normally distributed with mu and sigma square and we define a linear function on x, Y is a linear function of x then Y is also normally distributed.

So, Y is defined as a plus b x, it is a linear function of x then Y is also it can be shown very easily that Y is also normal distributed with the parameters a plus b mu and b square sigma square; that means, the mean of the random variable Y will be a plus b mu and the variance of random variable Y will be b square sigma square. Now we look at the cdf of x. So, f of x is given by minus infinity to plus infinity minus infinity to x this is the cdf

we are talking about f of dx which would be $\frac{1}{\sigma\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ and this is for x varying between minus infinity to plus infinity.

Now, if you have a sample and now you have estimated the you have obtained the sample estimate for μ and σ then you should be able to get the F of x , because the f of x is completely defined in that case, and therefore you should be able to get f of x , if we are able to integrate this between minus infinity to x . Also as your μ and σ change for different samples your pdf itself changes because it is a two parameter pdf therefore, as the parameter value change your pdf changes; and therefore you need to obtain this integral for a specified value of μ and σ , because the integration is not easily possible.

We adopt numerical integration and when we want to do numerical integration if you have to do it for every given μ and σ then it becomes cumbersome. So, we use this important result that if we define a function on random variable x a linear function on random variable x , then we know that y also follows normal distribution with parameters specified this specified like this. So, we should be able transform, if you have the random variable x you should be able to obtain a linear transform of x which also follows the normal distribution with parameters $a + b\mu$ and $b^2\sigma^2$, we use this result and generalize the integration for the normal distribution.

(Refer Slide Time: 48:45)

Normal Distribution

$$Z = \frac{X - \mu}{\sigma}$$

$$a = \frac{-\mu}{\sigma}, b = \frac{1}{\sigma}$$

$$Z \sim N\left[\frac{-\mu}{\sigma} + \frac{\mu}{\sigma} \cdot \frac{1}{\sigma^2} \times \sigma^2\right]$$

~ $N(0,1)$

pdf of z

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < +\infty$$



cdf of z

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-z^2/2} dz \quad -\infty < z <$$

-- Linear function

$$Y = a + bX$$

$$Y \sim N(a+b\mu, b^2\sigma^2)$$

So, look at this transformation z is equal to x minus μ by σ this is linear function and in our earlier notation from our earlier notation where we wrote y is equal to a plus b x , if we use that here a will be minus μ over σ this part and b is one over σ from, here. So, because y follows in particular case a normal distribution with a plus b μ and b square σ square and this is the mean of random variable y and this is the variance of random variable y .

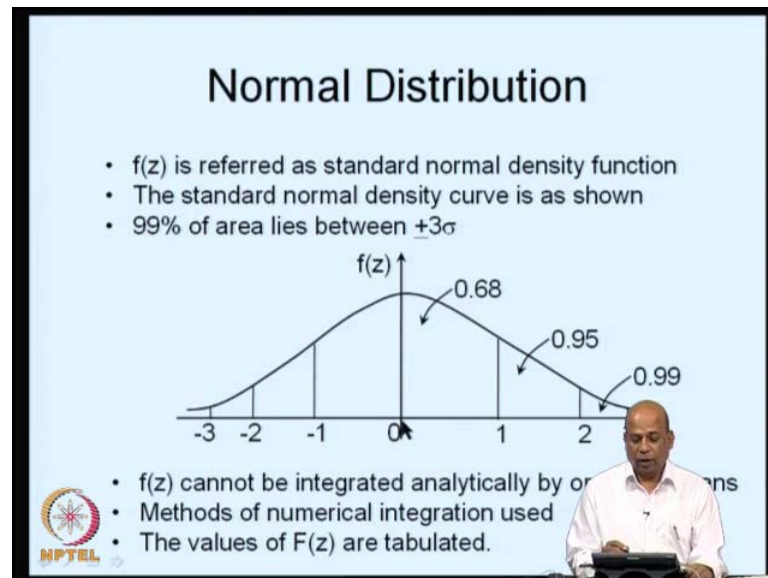
So, substituting this here z should follow normal distribution with a plus b μ . So, this is a plus b μ . So, you get this as 0 and 1 b square σ square as a variance. So, you get the parameters of the normal distribution that z follows as 0 and 1 and this becomes a extremely handy result because once we have this transformation z equal to x minus μ over σ z follows a normal distribution with 0 mean unit variable, and therefore we should be able to define the cdf of z and use the integration to obtain the associated probabilities on x .

So, once we define z we define the pdf of z , because z follows normal distribution we obtain the pdf of z as 1 over root 2 π σ is 1 therefore, we do not write σ here e to the power minus z minus μ x the whole square there and μ is 0 here for this therefore, you get e to the power z square by 2 and z varies between minus infinity to plus infinity because x varies between minus infinity to plus infinity. So, once we get the pdf of z you get the cdf of z as integration between to z 1 over root 2 π that is this f of z and you get this integral here minus infinity to z e to the power minus x square by 2 $d z$. Recall that your cdf of z what does it give it gives you that is you are talking about f of z this gives you probability that z is less than equal to z . So, for a specified value of z we should be able to get the probability of z being z being less equal to value if we can integrate this function e to the power minus z square by 2 with respect to z .

However again the integration of this is not easily possible with our usual routine method and therefore, we have to go for numerical integration, but the advantage here is that because it has a 0 mean and unit variance it does not have any parameters and therefore, for a specified value of z you can integrate this and then give the integral of values there is probability of z being lesser equal to z for various values of z and that is what is done using the numerical integration. So, what is we have did starting with the normal distribution we obtained a transformation on the random variable x we defined z is the now called as the standard normal deviant or standard normal variant. So, from x we

transform a variation z is equal to x minus μ over σ , and then see that z follows a normal distribution with mean as 0 and standard deviation as 1, let us look as how the standard normal distribution looks like. So, the $f(z)$ which is the pdf of z is referred as standard normal density function and if you plot the standard normal density function it look like this it is perfectly symmetrical about x is equal to about z is equal to 0.

(Refer Slide Time: 53:06)



Then another interesting property here is, if you take one standard deviation on either side the standard deviation of z is 1 therefore, if you take 1 standard deviation here and 1. Standard deviation here the area contained within that is about 68 percent about 68 percent pdf is contained in 1 standard deviation around the mean which is 0 and you take 2 standard deviations that is between minus z is equal to minus 2 to z is equal to plus 2 about 95 percent of area containing that. So, this area between z is equal to minus 2 to z is equal to plus 2 contains 95 percent of the area and similarly three standard deviation, if you take minus 3 to plus 3 that is z is equal to minus 3 to z is equal to plus 3 about 99 percent of the area of about this pdf is contained within this region.

Now, this has an important notation or implication that when you are dealing with some normal processes not really the extreme processes like this that is you are not concerned about this what is happening to the tail of the distribution or on the extreme right of this distribution about 99 percent of the values lie within plus or minus 3 about standard deviations. Now, because as I said $f(z)$ cannot integrate analytically by ordinary means we

use numerical integrations and then tabulate them for various values of z we tabulate the area under the standard normal curve or the standard normal density function and then given a value of x for which we are interested in obtaining the probability. Let say that probability of x is less than or equal to a given value of x you are interested in. You use the transformation z is equal to x minus μ over σ and then start talking about the probabilities on z rather than probabilities on x using the tables that we generate using numeric integration. So, with this now we will pause for today and we will continue the discussion in the next class.

So, what is it we did that today we started with the definition of the moments we take the moments about the origin define the moments of the origin first. So, the first moment of the origin we call it as expected value, and then once we define the expected value we start taking moments about the expected value and define the higher order moments about the expected value itself.

We introduced measures of central tendency as the mean or the expected value itself and the mode and the median then we went on to see the measure of a measure of dispersion, which is the variance which we defined as second moment about the mean, and then we also introduced the measures of symmetry which are based on the which is based on the third moment about the mean, and we also obtain the associated sample estimate then we the measures of peakedness which is a kurtosis and kurtosis is defined based on the fourth moment about the mean. Then we look it at the most commonly used distribution which is a normal distribution we saw that the normal distribution is perfectly symmetrical and therefore, the coefficient of skewness is 0, and the kurtosis is 3 and then normal distribution important property that a linear function define on the random variable following the normal distribution also follows normal distribution.

And we use this result this important result to obtain what is called the standard normal density function, and we define the cdf of the standard normal density function or standard normal variable. And now we are move on to tabulating the cdf values for the standard normal variable. So, we will continue the discussion in the next class, and we will solve some examples using the normal distribution. And we will acquaint ourselves with the use of the standard normal table. Thank you very much for your attention.