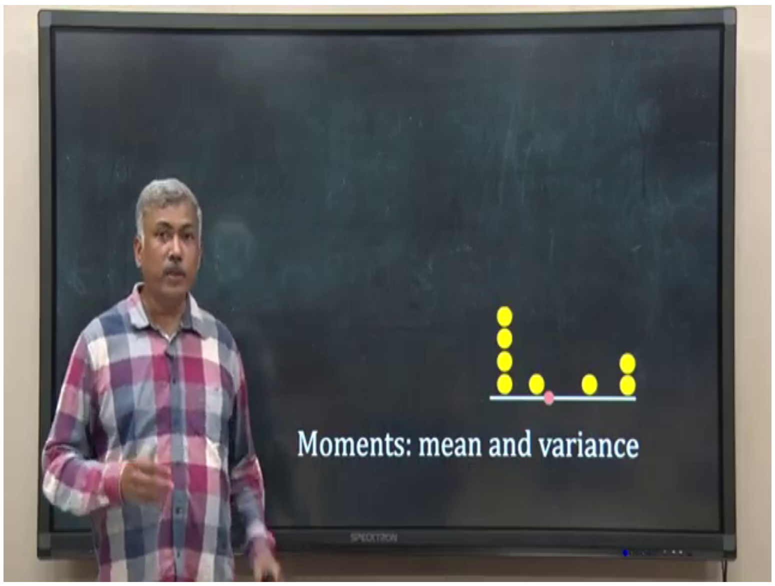
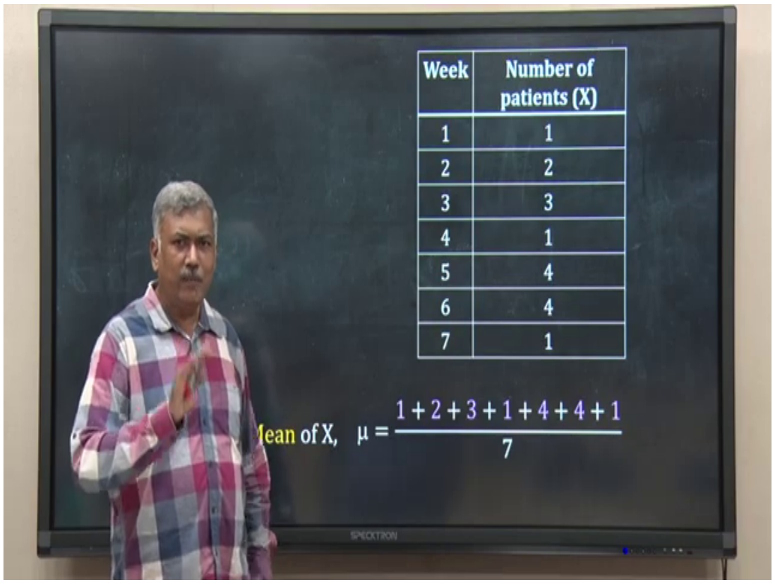
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture: 4
Moments: Mean and Variance

(Refer Slide Time: 00:31)



Welcome back. In this lecture, we will discuss about the concept of moments in probability theory, particularly I will discuss two very important moments, mean and variance as those will be frequently used in our data analysis. So, suppose I am collecting data from a cancer hospital and I have shown some data here.

(Refer Slide Time: 00:52)

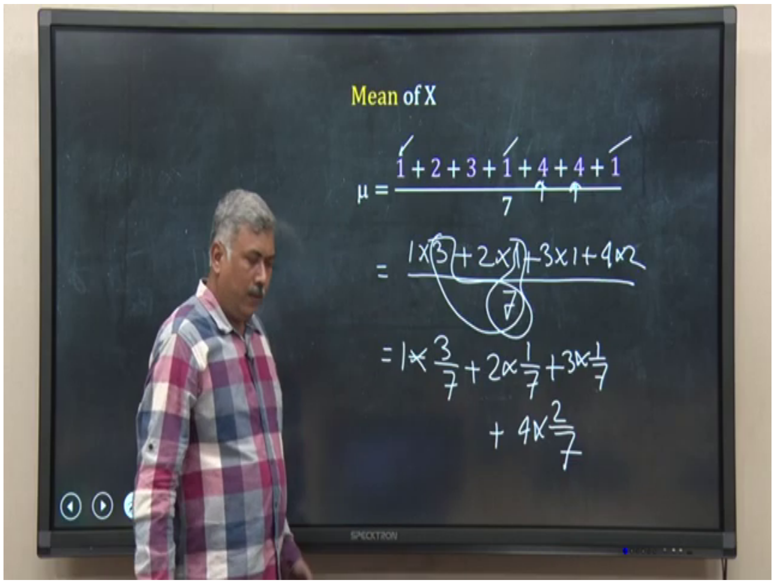


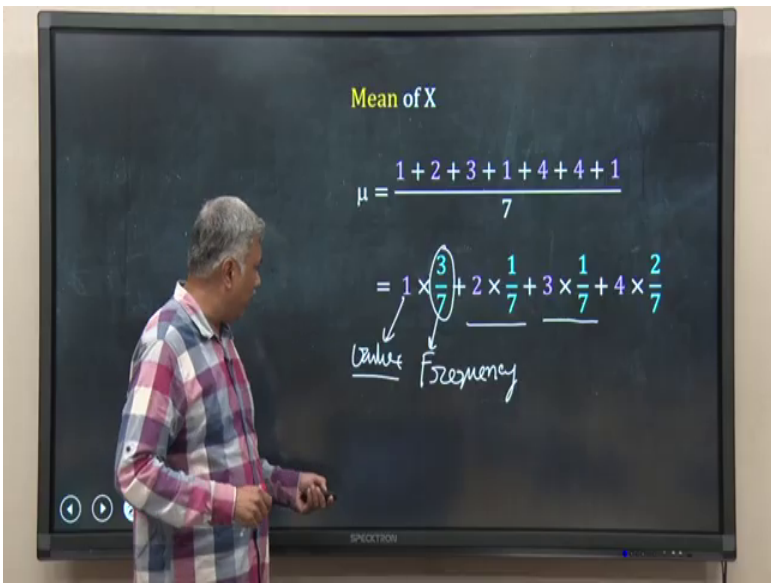
So, I have collected the number of patients, particularly suppose throat cancer patient who are getting admitted to that hospital per week, the number of patients. So, first week we have 1, third we have 3 people and 6th week we have 4 people. So, in this way suppose for 7 consecutive week, I have collected some data, number of throat cancer patient admitted in that week in that hospital.

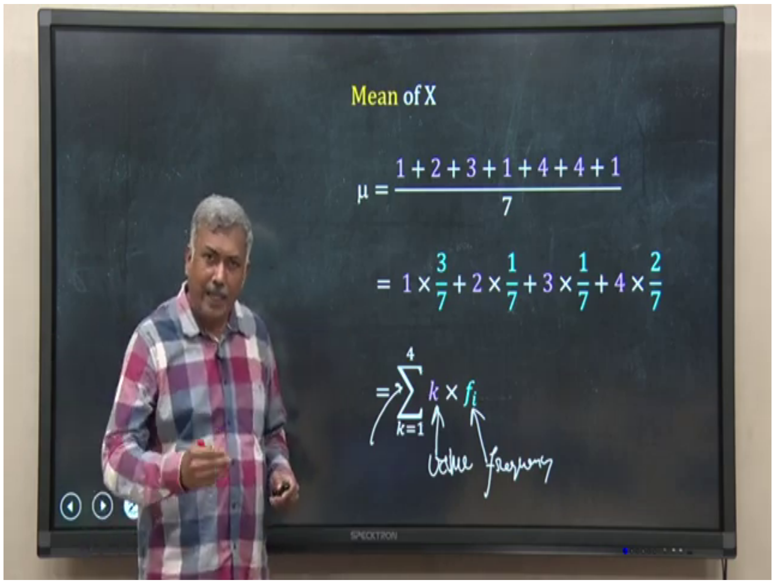
Now, if you have to consolidate the data, what you will do your common sense say, I will not report this whole length of table what I will do, I will report the average number of patient, throat cancer patient admitted to that hospital per week. So, what you will do, you will calculate what we call the mean, we will calculate the average.

So, calculating average is very easy, what you will do you will sum those all those term and divide by number of sample that is 7 and you will get the average value, average number or mean number of patients who are getting admitted to the hospital. Now, I want to do the same mean, same average calculation, but in a bit different way, let us see that.

(Refer Slide Time: 02:00)







So, I have written,

$$\frac{1+2+3+1+4+4+1}{7}$$

That is what you will do if you have to calculate the average. Now, let us count how many ones are there. So, I have one 1 here, second 1 here and the third 1 here. So, I have three 1s, I have two 4s, this one, this one and 2 and 3 is coming only once.

So, I can rewrite this one something like this, I have

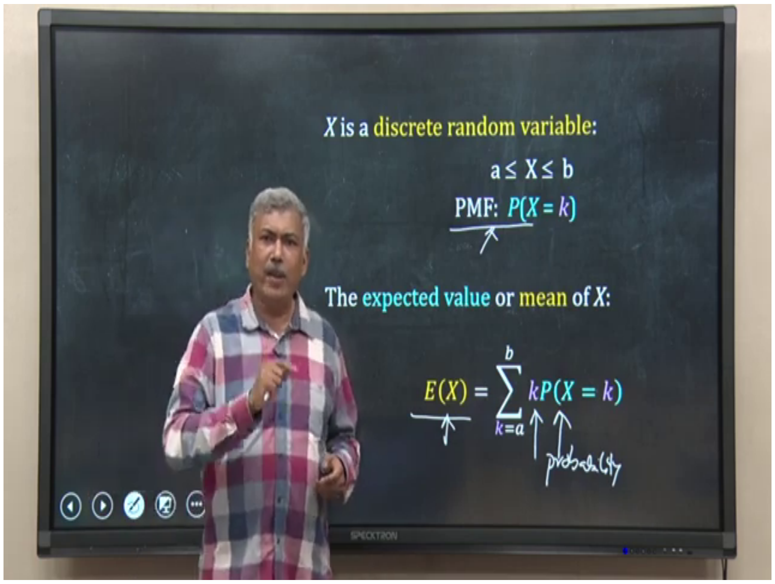
1*(3/7) + 2*(1/7) + 3*(1/7) + 4*(2/7)

I can write it something like this, 1 into 3 by 7, plus 2 into 1 by 7, plus 3 into 1 by 7, plus 4 into 2 by 7. I can write that.

So, that is what I have done here. Now, notice what are these, what is (3/7), this is frequency. How many time 1 has appeared in my data, and what is the value, this is the number value. So, each of these terms is value multiplied by frequency, value multiplied by corresponding frequency.

So, I can generalize this using mathematical symbol, what I can say what I am doing is I am multiplying the value, a particular value, the value here is having from 1 to 4, 1, 2, 3, 4 and its corresponding frequency. And if I sum them together, this symbol is for summation, if I sum them all together, what do I get? I get the average or mean, the humble mean that we have learned in school. So, now, let us make it generalize.

(Refer Slide Time: 04:16)



Suppose I have a random number X, which is discrete random number, a random variable, which varies from a to b, and I know the PMF Probability Mass Function I know, it is given here P(X) = k. So, this is a function, probability mass function which gives me the probability of a particular value of X.

So, then the expected value or mean, remember this is a technical way of saying mean, it is called expected value. So, the expected value or mean of X is E(X), this is a symbol we write because it is expected value you write E and in bracket the random variable. So, E(X) expected value, or mean of X is equal to

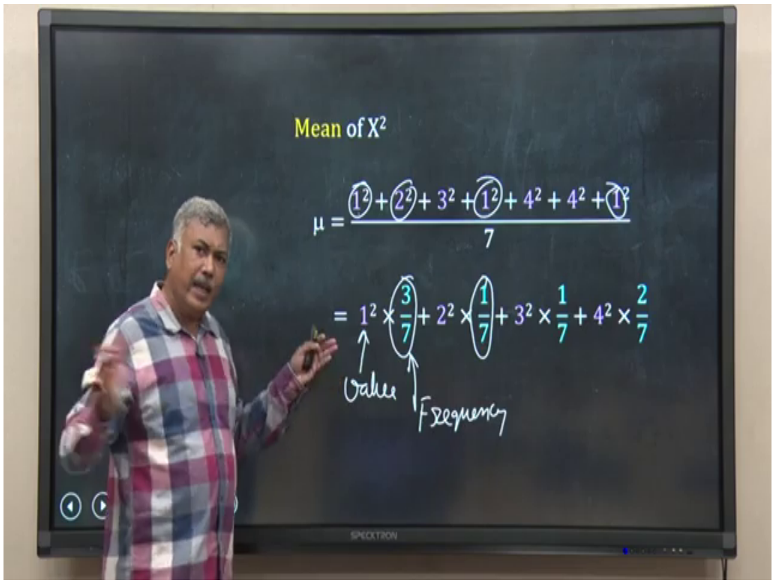
$$\Sigma\ k*P(X=k)$$

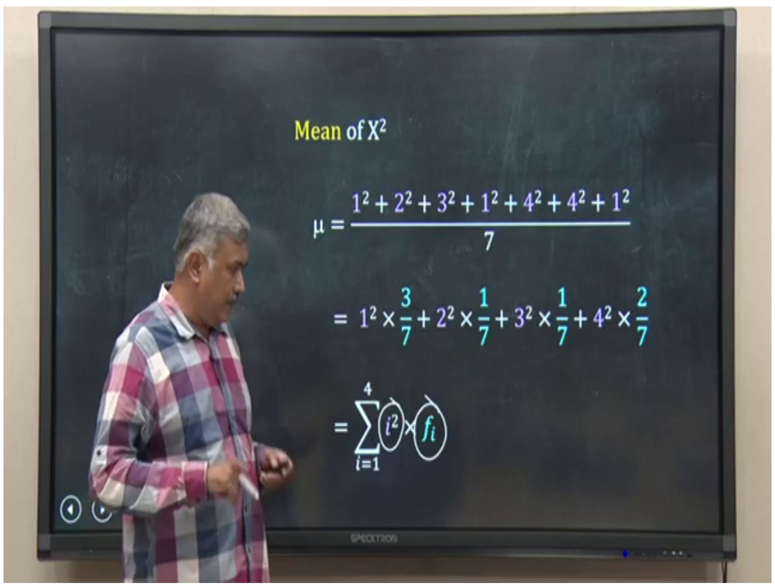
summation of a particular value of X into the probability.

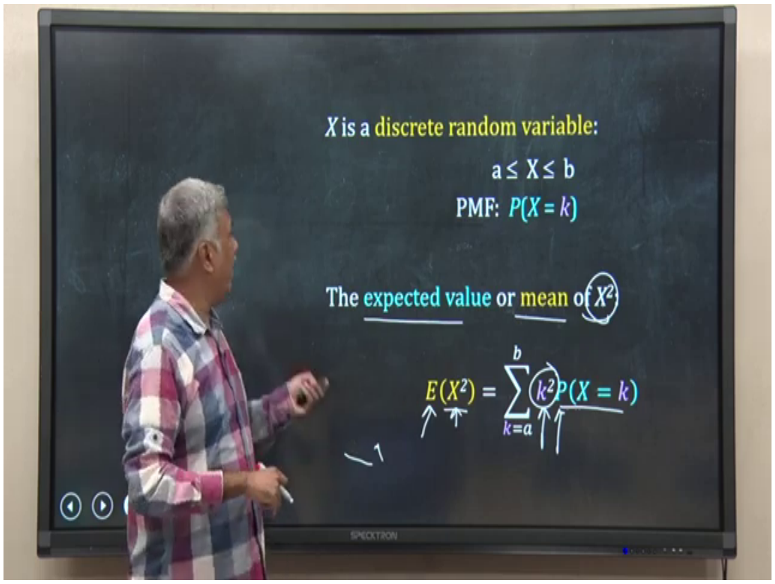
Where I am getting the probability from, I am getting the probability from the probability mass function. And if you remember, in the example of patient data, I had frequency there. And usually, if I have a large data set, we consider that for a large sample size, frequency is equivalent to probability, so we have just mapped that, we have replaced frequency by probability.

So, what I have got, I have got summation of values of X into their corresponding probability. And that is my mean, or expected value of a random variable, which is a discrete random variable. Now, I want to use the same technique, but I want to calculate the expected value or mean of not X, X square, square of X.

(Refer Slide Time: 05:49)







So, what you will do, simple, you will again use a school level idea of mean, as simple as that. So, you will take 1^2, 2^2, 3^2, 1^2, 4^2 like that, and you will sum them and divide by 7, because 7 is your sample. I can again do the rearrangement here, I had 1 square appear 3 times, see here and here.

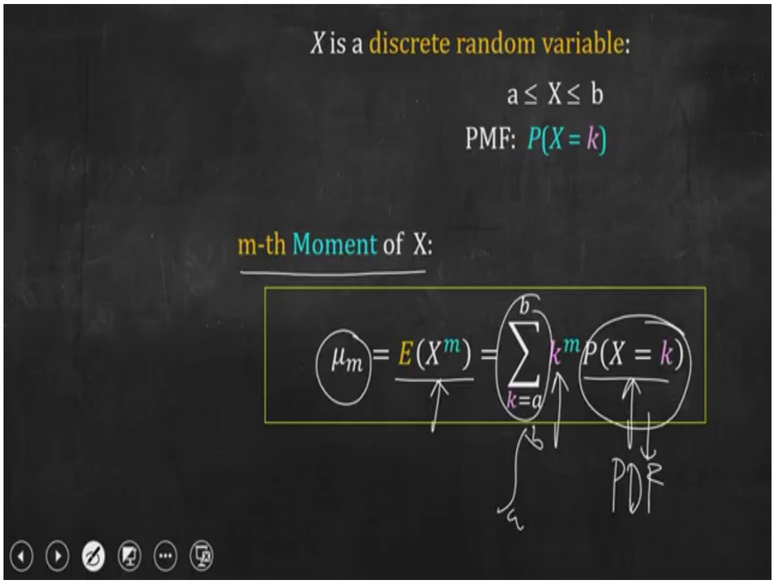
So, I have (3/7), that is the frequency, how many times 2 square has appeared, only once. So, I have frequency (1/7). Similarly, 3 square has appeared only once. So, its frequency is (1/7), whereas 4 has appeared twice. So, 4 square has appeared twice, so its frequency is (2/7). So, I have just multiplied the value.

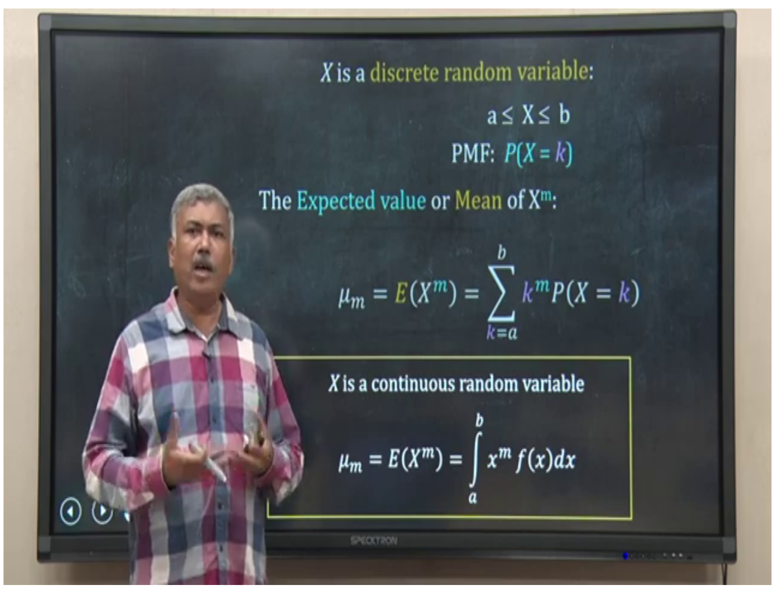
Again, just like the way we did for mean, and the earlier case, and its frequency. I am multiplying them and then summing up. So, using the mathematical symbol, I can write, it is summation of the value into its frequency, and then sum them together. So, then again, just like the mean, I can do a generalization, X is my discrete random variable, which varies from a to b.

And I know the PMF Probability Mass Function, I know, then the expected value, or mean of X square in this case, remember, we are dealing with X square, not X, of X squared will be $E(X^2)$ expected value of X square will be equal to a value square into the probability, these are probability. And this is a square of that particular value. So, I have generalized it. So, I have generalized it for a discrete random variable for mean of X.

Or E(X), mean of X squared or $E(X^2)$ Why do not I go further up? I mean, why do not I go to X^3, X^4, and something like X^m, generalizing it broadly. So, what we will get.

(Refer Slide Time: 08:00)





So, again, X is a discrete random variable varying from a to b, and I know the PMF, suppose I know it, then I can write that the $E(X^m)$, we are done with X to the power 1 that is X, we have done with X^2. Now, I am saying $E(X^m)$ will be equal to, using the same logic, a value of that particular variable X^m into the probability. here the probability remains same, only the term before probabilities is raised to m. And then you sum them all together from a to b, k will vary from a to b.

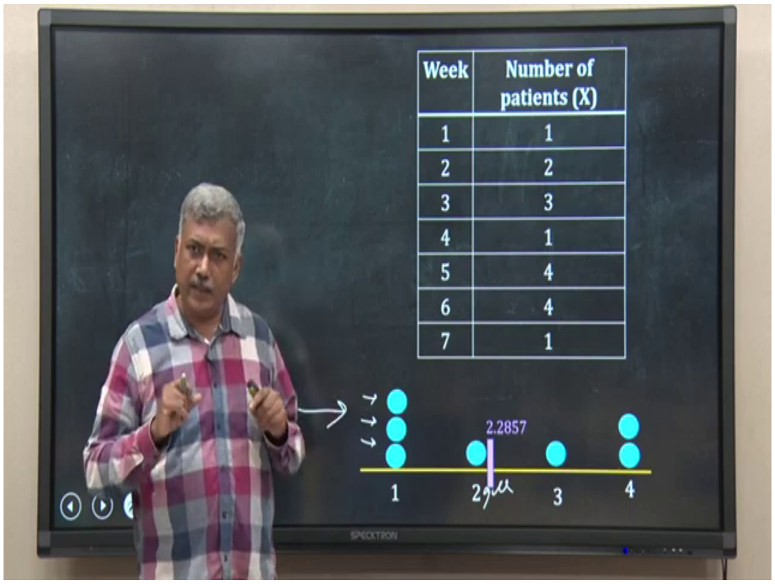
So, this is the expected value of X to the power m, it is most often written as μ_m. So, this is the generalized definition of moment in probability theory, you may have heard that in many lectures or in textbook or some literature you have heard people are talking of a first moment.

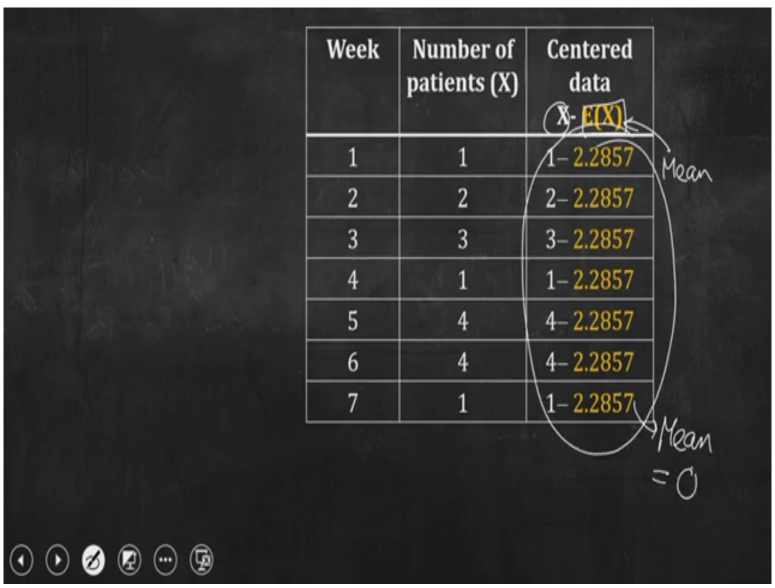
Third moment, something like that, this is the moment. mth moment will be $E(X^m)$ and this will be the formula. So, in the first moment, m = 1, in second moment m = 2. So, if you have paid attention to our concept of discrete random variable and continuous random variable, you must have noticed what we are dealing till now is all discrete random variable.

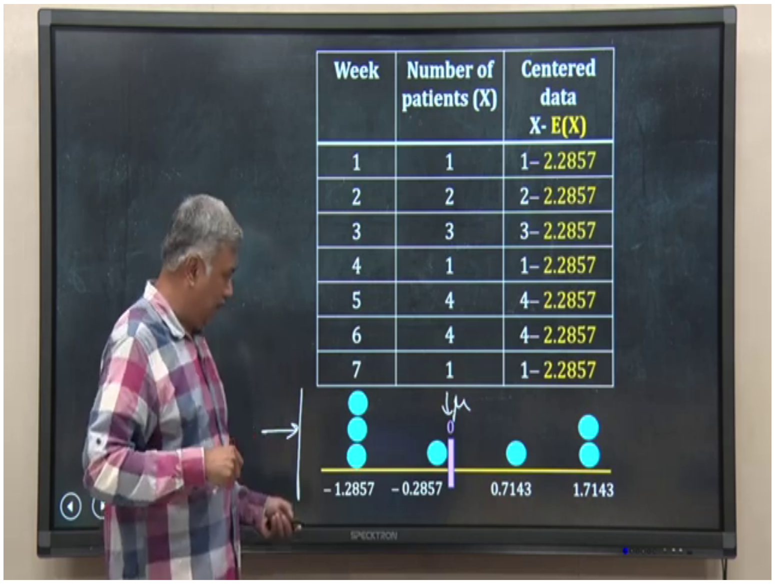
That is why we have PMF Probability Mass Function. And you may be wondering, what should I do if I have a continuous random variable X, is the continuous random variable? I should have a definition of moment for that also, yes, nothing will change much it is very simple, what it will happen I will replace this P probability mass function by the PDF probability density function and the summation will be replaced by integration sign.

So, it will be integration from a to b. So, that is what you will do. That is what I have shown here, we will not go much in detail of this, because going into detail of the summation, integration is not required for this course, just we have to understand what is the concept of moment and how it emerged, it comes out simply from the idea of average or mean that we all have learned in school that.

(Refer Slide Time: 10:20)







So, I have the same data, what I have done, I have plotted that on a number line here and each circle, each of these circles represent each of those result count. So, 1 has appeared 3 times, whereas, 4 has 2 times and this pink one is my mean μ or E(X). Now, when you consolidate a data in a large table like this, it may have 700 data points or something like that.

So, you not only report the mean, but at the same time, you also want to know the dispersion of the data, just telling me that the average number of patient getting admitted, throat cancer patient admitted in the hospital per week is 2.2857 does not really give me the clear picture, because I also want to know how much is the variation from week to week.

And this dispersion issue, how far it can go, how low it can be, how big it can be, this idea these variations remember, is always with respect to the mean behavior. I want to know how much it will vary from this mean, how much it will vary from this particular mean value. Now, if you suppose want to compare the dispersion or variation of two different data sets, I have one for suppose throat cancer this one and I have another data set for breast cancer patient.

Now, obviously the means will be different possibly and their dispersion will be also different. Now, if I have to compare the variation of these two different data set around their mean, now, if their means are different, then it becomes a bit complicated that does not make much sense isn't it, because their means are different, what I want that I have to do something so that the data set mean become same.

The simplest way to do that is that universally when I will get this data, I will make, do something some mathematical transformation I will do so that the mean will become 0, I will do that for every data set. So, that is what called centering of data and it is used for many purposes, I am just showing for one example here, for one particular reason, because I want to know the dispersion of the data around the mean.

So, I am doing centering what I am doing, what I have done here is, I have taken the value and subtracted the mean of that because this is my mean. So, if you remember just in a previous slide, I have shown the mean is 2.2857. So, I am subtracting that value from each of these values. So, I have centered the data, why I have centered the data, I have now subtracted the value of mean from each of them and then what happens.

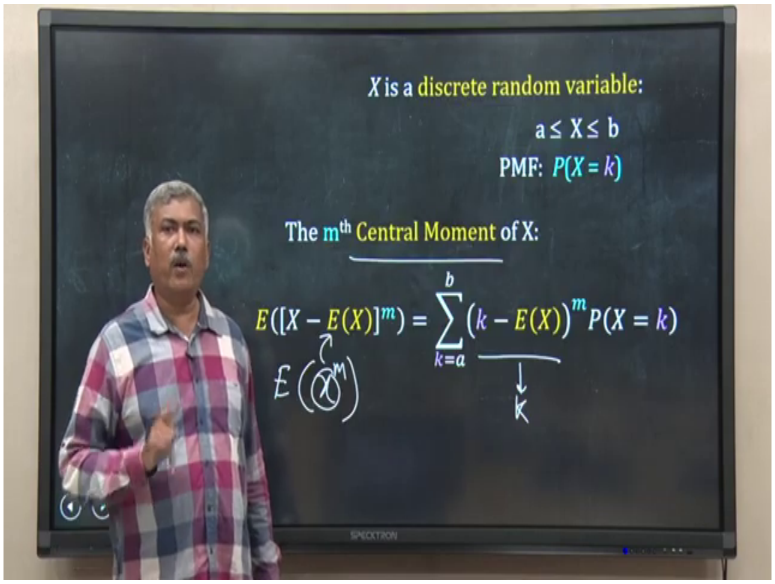
Actually, if you calculate the mean of these values, if you calculate the mean of all this centered data, you will find the mean will be 0, you can try that, just a caution here. The mean is actually not 2.2857 there are some other digits after that what I have done just to accommodate in this table in this space I have truncated. So, if you try that in a Excel sheet or somewhere or in a calculator, you will find that if I use 2.2857 as the mean value, you will not actually get a 0 as mean.

But it will be a very small number. So, you have to accept that is equivalent to 0, otherwise, you can take all the digit and actually the mean which should come to 0. So, what I have done, I have centered the data try to understand why I am saying centered look at this data now. Now, the mean is 0, this is my mean now, and the all the data is dispersed around the 0.

You may have a breast cancer data set, you may have a liver cancer data set, you may have a glioma data set all cases you can do this type of centering with respect to their respective means and then mean of all this data set will become 0. Now, you can ask, tell me what is the dispersion around the mean? So, the issue of mean, variation in mean is gone.

Now, I am safe, I can talk of variation. So, what we will do now we'll calculate the moments, we learned the idea of moment. So, we will calculate the moments on the centered data, not on X, we will do the moments, apply the concept of moments on X minus E of X. So, how it will look?

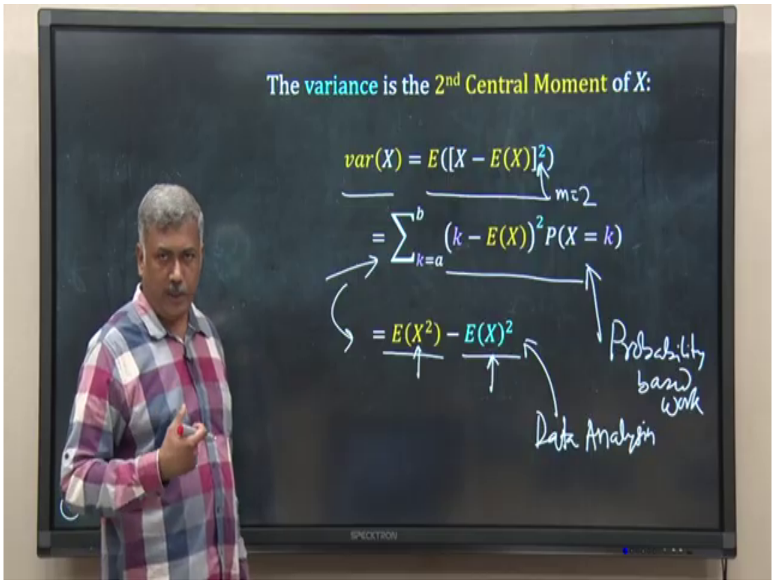
(Refer Slide Time: 14:42)

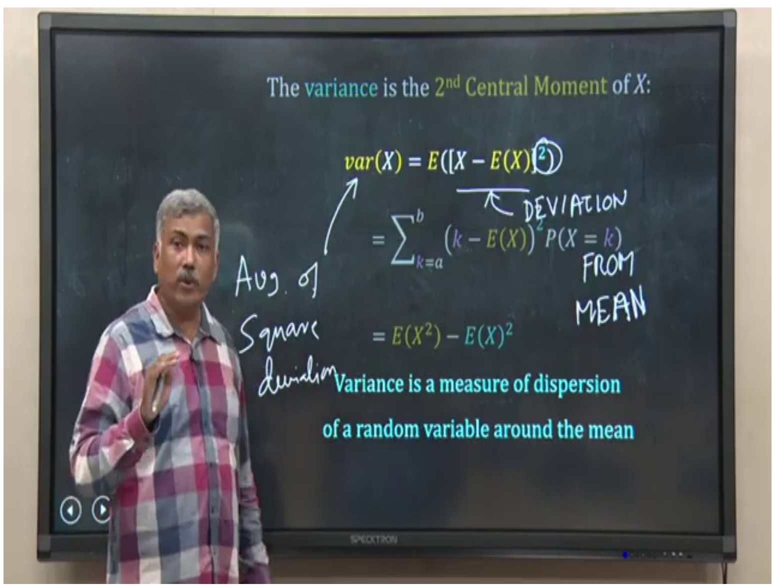


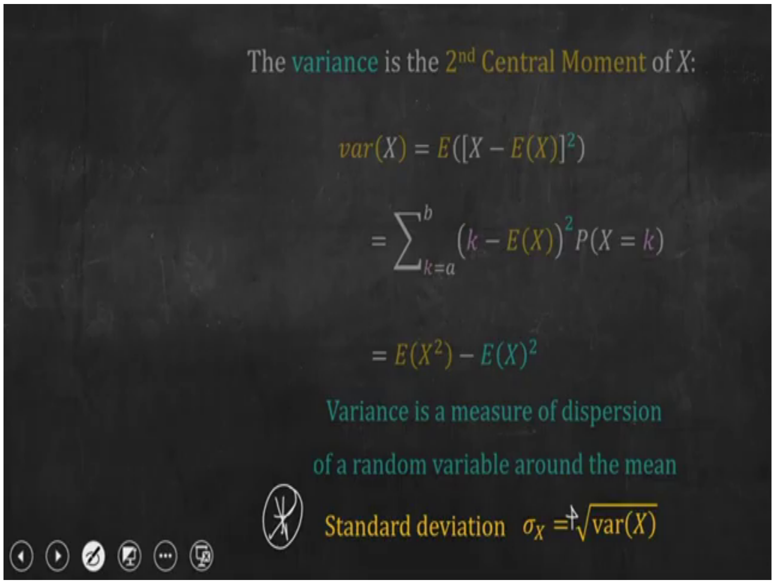
The same thing everything will remain same. In the earlier case I was writing $E(X^m)$. Now, X is, I do not have X, I have X – E(X) because there is a centered data. So, I take $E(X - E(X))^m$, because it is mth moment, that will be equal to earlier I had k here in the earlier when I was talking a moment I had k here.

So, now in this case I have k – E(X) the mean because it is the centered data in to the power m because it is mth moment into the probability coming from the PMF. So, this is called central moment, because you are working on centered data. Now, one central moment is very important for our data analysis course that is called the variance.

(Refer Slide Time: 15:38)







Variance by definition is the second central moment, you must have learned variance earlier in some other course. So, variance of X a random variable is the second central moment, see here m = 2 of what, it is the moment, we are taking for the centered data X – E(X) that is why it is a central moment, it is second central moment and you can use the formula of moment and the definition of moment it will be this one as simple as that power will be 2.

And you can do some sort of simple algebra to it can be shown is that this one is the variance of X is equal to, expectation of X^2, you are squaring the data and calculating the expectation, we did it a few slides back we did that, minus square of the expectation of that X. This you can prove algebraically although you do not need to prove that for this course.

But we have to remember these two has define uses. For example, for data analysis work for large data analysis work, we use this formulation.

$$\text{Variance} = E(X^2) - (E(X))^2$$

This formulation becomes much easier you can write if you are doing, writing a code to do data analysis, this formulation becomes much easier to do. Whereas, if you are doing some probability based generalized work this one is more useful.

So, we have reached variance which you must have encountered earlier and what we are measuring here, that we have to understand the physical meaning, we have got the, it is a second central moment, but what actually it is measuring, what is the physical meaning of variance? Let us look into that. So, the variance is actually a measure of dispersion of a random number around its mean.

Why I am saying that, because see what we have here, this is what X – E(X), so, that is mean of X is subtracted from X. So, this is deviation, this is nothing but deviation from mean. So, X – E(X) is a deviation of a number from its mean, of a value from its mean. Now, that deviation can be positive, that can be negative. Just in the previous example, you have shown, seen when I am centering the data, mean is becoming 0, which was earlier a positive one, it is becoming a negative value.

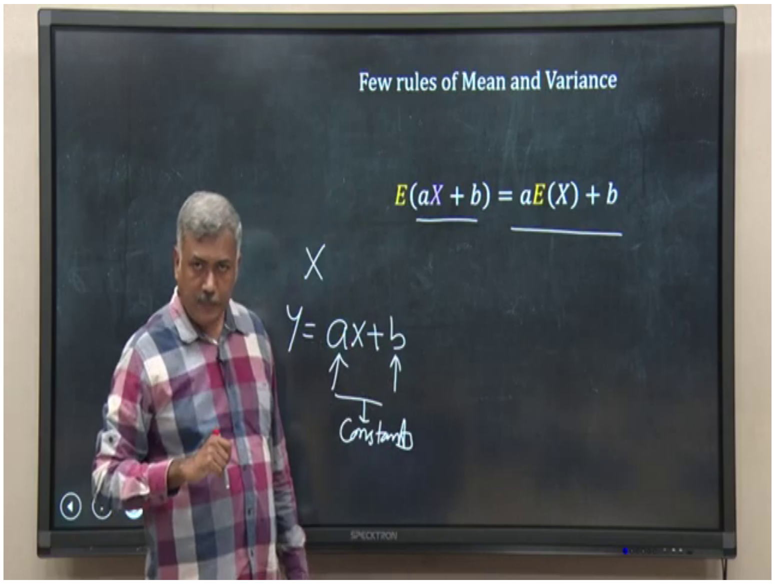
So, sometime it will be negative, sometime it will be positive. So, now if I now try to calculate the mean of that deviation, then as some are negative, some are positive, my mean will be 0, that will not make any sense. So, what I am doing, I am taking the square of it. So, I am taking the square of deviation from the mean. So, once you square negative and positive, all will be positive and then you are taking the average of that square deviation.

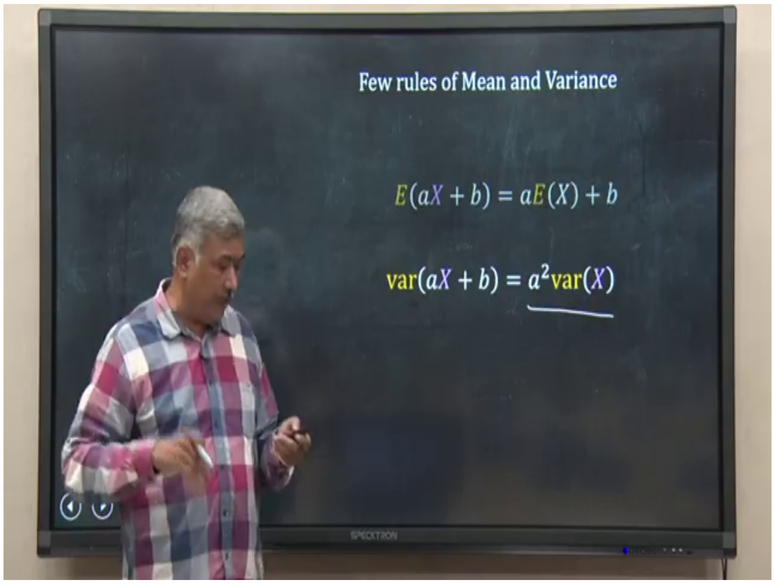
So, what it is? Variance in a way is nothing but average of square deviation, in a way, because expectation is averaging. So, you are taking the mean of squared deviation. So, for some data points that deviation is very large, some data point that deviation is very low, I am talking of deviation from mean and you are averaging those deviations square. So, you are getting variance.

And as it is average or expected value of a square thing, remember square things are always positive, variance has to be positive always. Usually, we do not report variance. When you do data analysis, we do not usually report variance, what you report is usually standard deviation and that is very easy to understand, standard deviation is nothing but the square root of the variance.

Now, here one point we have to remember, see if I take the square root of 4, you get two answers, +2, -2. Now, which one I should take here, should I take the negative value also for root of square root of variance? No, by convention, we only take the positive value. So, standard deviation can also not be negative it will be positive by convention because it is a measure of distance, it is a measure of distance from the mean. So, that measure is the absolute term, so it is always positive and it is a positive value of the square root of the values. That is what we do.

(Refer Slide Time: 20:44)





So, suppose I am measuring X, X is a random variable. For example, suppose X is absorbance in, of a suppose, you are doing a calorimetric measurement of some molecule like protein molecule solution. So, X is the absorbance. Now, from that absorbance, you have to get the value of concentration of the protein solution.

So, what you do, you have a standard curve, and that from standard curve, you get some constant term, you actually use that constant term and multiply the absorbance value to know the concentration of your sample, that is what you do. So, what you are doing, you are measuring X but you are reporting something constant multiplied with X.

So, suppose I have a random variable, which I am measuring X, but from that, I am creating another new variable which is nothing but a*X+b, where a and b both are constant, they are not random variable, they are constants. For example, a dilution factor, it is the constant. The sampling that you are doing is random, but the dilution factor will remain constant.

Or suppose as the example I gave the protein solutions absorbance, when you convert from absorbance to concentration you multiply with a constant term. So, if this is the situation then, if I know E(X) what will be E(a*X+b). So, it can be proved and we do not need to go to the proof just we have to remember this rule that

$$E(a*X+b) = a*E(X) + b$$

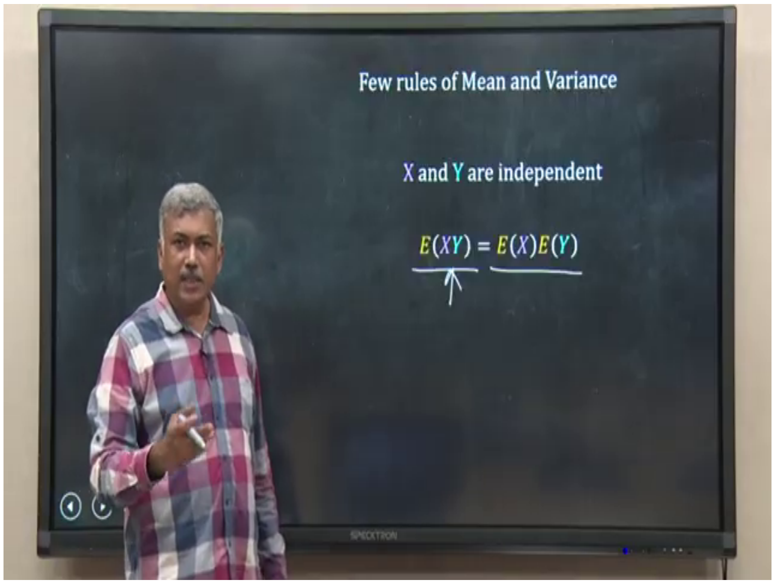
Where a and b are constant, X is the random variable, it does not matter whether X is a continuous random variable or discrete one in generalize it is a random variable. Now, what will be the variance in this case, variance is very interesting,

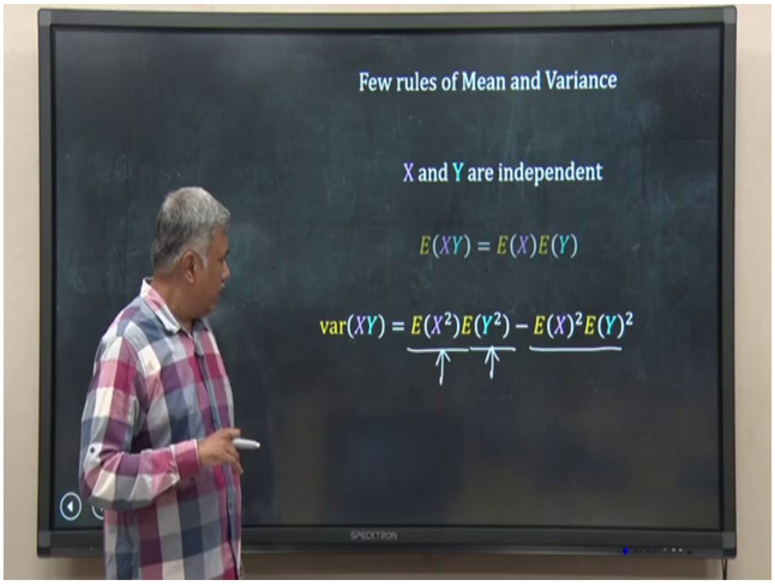
$$Var(a*X+b) = a^2 * var(X)$$

Remember b does not appear there. Because b is a constant and there cannot be any variance of a constant term, it does not vary. So, that is why there is much be 0, which I have not written. So, it will be a square into variance of X. Now, this is for, something I am multiplying and adding with a random number X, but suppose I have two random numbers or suppose in other way.

I have two measured random variable, two parameter I am measuring in the experiment, and when we report the data, we actually multiply these two, and imagine that these two random variables that we are measuring are independent of each other that means, the probability of one does not affect the probability of the other one.

(Refer Slide Time: 23:33)





So, in that case what will happen, so, for here X and Y are the independent random variable, again it does not matter whether it is discrete or continuous, then the expectation of their product,

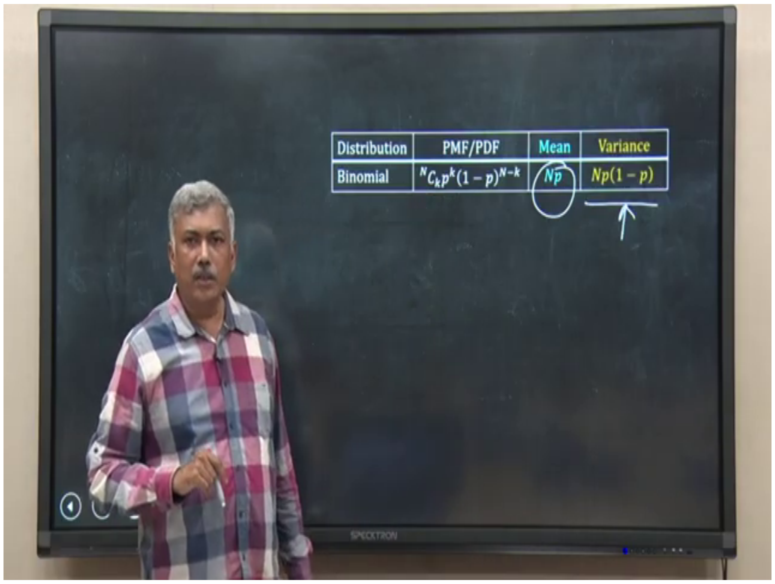
$$E(XY) = E(X) * E(Y)$$

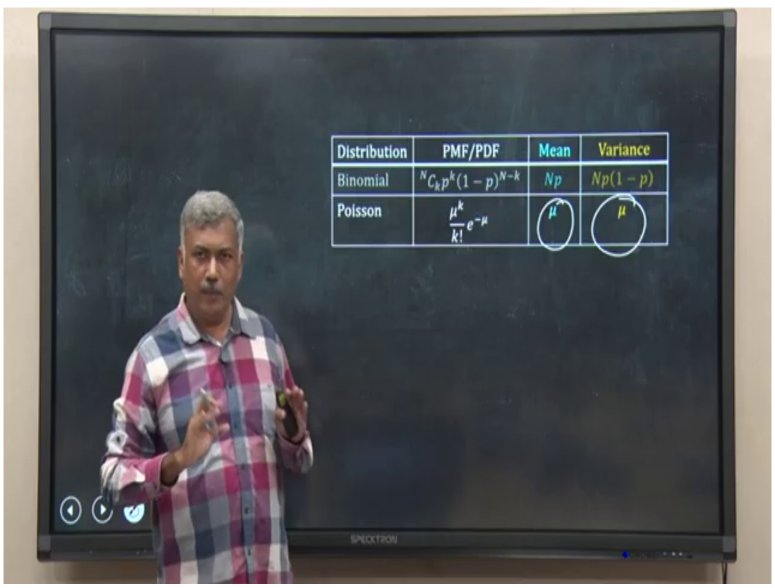
E(X) is the mean of X, E(Y) is the mean of Y. What will happen for the variance,

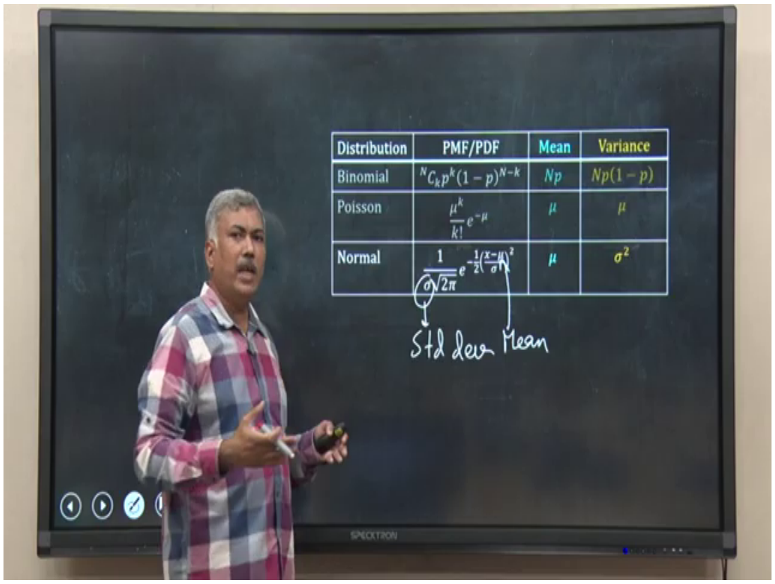
$$Var(XY) = E(X^2)\, E(Y^2) - E(X)^2\, E(Y)^2$$

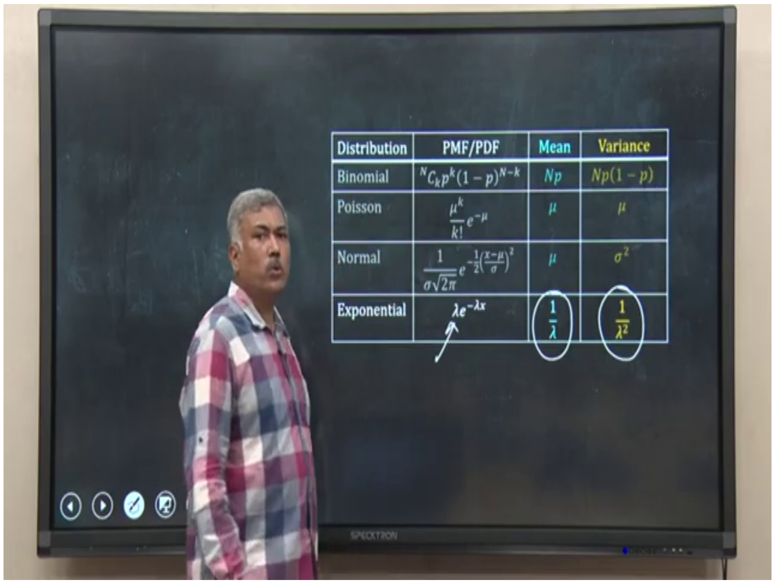
So, now, let us move to something else.

(Refer Slide Time: 24:14)









Which is connected to the same concept of moment, if you remember I said I want to understand the mean and variance in terms of moments because, one it helps us to combine different data and calculate the mean and variance for them using those rules that I discussed just now. And the other reason is that, if I know the PMF for the PDF, or in other words.

I know the, I am doing some generalized calculation, I know the probability distribution of the variable and then I want to know what is the mean and variance of that. In many statistical analyses and data analysis, the model that you build you already imagine that, there is a hidden poisson distribution or hidden normal distribution. So, what will be their mean and variance of those distributions?

Those can be calculated theoretically simply by algebraically, for each of the given PDF or PMF. And I have listed some of those, which are very common usually in biological data analysis, obviously, the binomial will come first. So, in binomial remember what we are doing, we have N number of trial and p is the probability of success, probability of having mutation, probability of death, probability of getting 1, probability of getting head something like that.

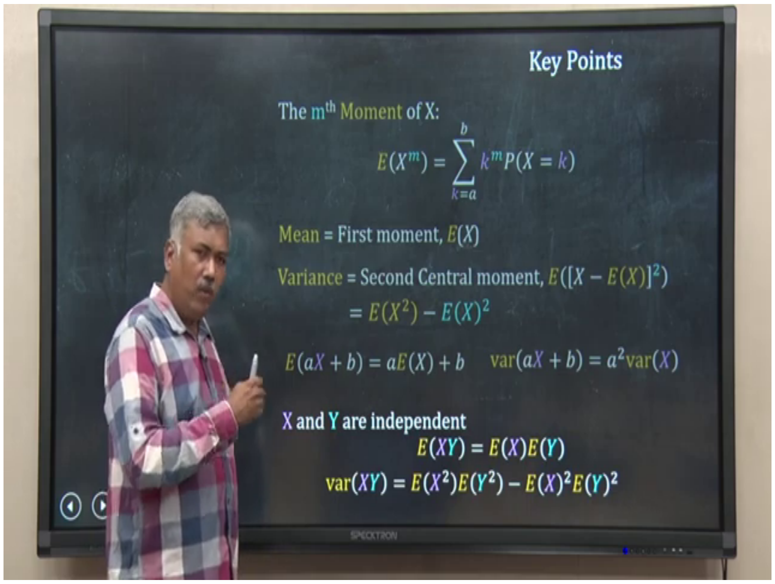
So, then the mean, you can prove, but you do not need to prove for this course, is will be N*p and the variance you can show will be N*p*(1-p). Another important discrete random variable that we have discussed was poisson distribution and poisson distribution has something very funny.

So, in poisson distribution, remember we had talking about some event, which is happening independently with a constant rate. And we are trying to find out the number of events happening in a particular interval of time and space. And in this case, the PMF has one parameter μ, the mean.

So, it can be shown that for a poisson distribution, the mean and the variance are both same, that is the μ the mean. Now, come to one of the most commonest continuous distribution that will encounter in any data analysis is normal distribution. And if you remember normal distributions PDF is defined by 1, this is the standard deviation and μ is the mean.

So, in any way your PDF itself is defined by these two parameters, standard deviation and mean. So, your mean and variance will be μ and σ^2, the standard deviation square. Now, take another one, another continuous distribution, exponential and if you remember exponential distribution has parameter λ. So, it can be shown that the mean of these exponential distribution will be $(1/\lambda)$ and the variance will be $(1/\lambda^2)$. So, that is all for this lecture, let me jot down what we have learned in this.

(Refer Slide Time: 27:20)



The first thing we have learned we have learned the idea of moment and we are trying to learn it because this has utility in different derivations and calculations, where probability theory is used. So, the mth moment of a random variable X, where the formula that I have shown it for discrete random variable.

But you can have equivalent one for a continuous one will be is equal to, the $\Sigma k^m * P(X=k)$, k is a particular value or realization of that random variable. Now, two moments are very important for our work. One is mean, the humble mean that we learned very early in our school or the average recall is nothing but the first moment of a random variable or measurements that you are making.

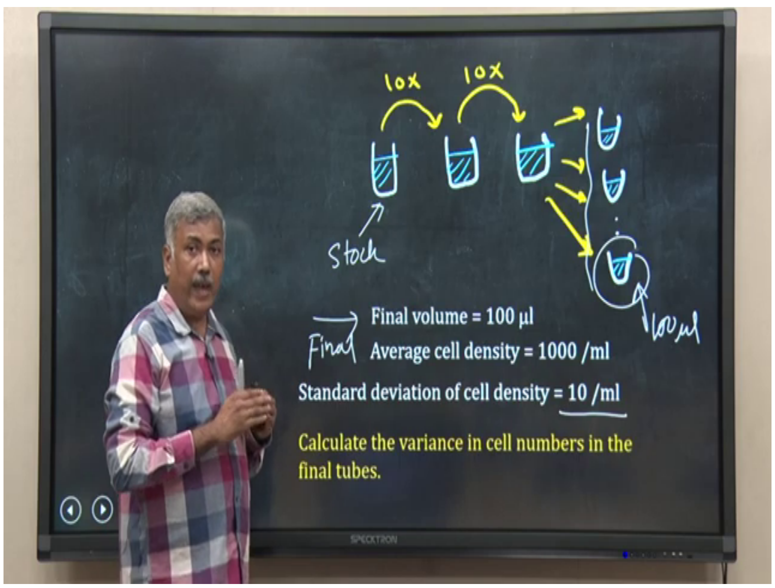
And the variance is the second central moment, it is central, because I have centered the data with respect to what, with respect to the mean. So, X – E(X), that is E(X) is the mean. So, and then we are taking the square of that, and then we are taking the mean of that square, and we can actually calculate it easily using these relationship is $E(X^2) - E(X)^2$.

So, this is the variance and then, we have learned certain rules, which can be derived from the definition of moments that if I have a random variable X, and if I take aX+b, then its expectation of this new variable will be a*E(X) + b and the variance will be a^2 var(X).

Similarly, if X and Y are independent, two random variable independent, then E(XY) will be E(X)*E(Y) whereas, the var(XY) will be $E(X^2)\,E(Y^2) - E(X)^2\,E(Y)^2$.

So, these are the generalized thing we have learned, I have not discussed how to calculate mean and variance in this particular lecture. We will have a separate lecture where we will discuss about not only about how to calculate variance from a data, but also, we will discuss our covariance. We will see, I will see you there in that lecture. But before I leave, let me leave you with a problem to solve.

(Refer Slide Time: 29:36)



Suppose, I have a cell suspension, the first you I have shown, this is suppose my stock cell suspension. And then I have done it serial dilution, that is what I have shown, x^2 10-time serial dilution twice and then from the serial diluted thing, I have sampled the cell suspension. Suppose 9, 10 different tubes.

And what we have measured, the data that is given to you is that the final volume is 100 microliters, that means, in each of these I have 100 microliters, 100 microliters of cell suspension and the average cell density in those final things, although it is not written here, average, this is the final in the final tubes those last left-hand side view.

So, the final average cell density is 1000 cells per ml and we have also measured the standard deviation, we have multiple 9, 10 tubes. So, we have checked what is the cell density in each of them, from there we have calculate the average or the mean cell density and also calculate the standard deviation. I have given you the standard deviation that is 10 per ml, 10 cell per ml.

Now, I want you to calculate the variance in cell number, remember the data given to you is for cell density, cell per ml, so I am asking you to calculate the variance in cell numbers in the final tubes meaning these tubes, these tubes, so, you have to calculate the variance from this given data, variance in cell number nothing else. So, if you have followed this lecture, you should be very easy to solve this one. So, try this one and join us for the next lecture. Till then happy learning.