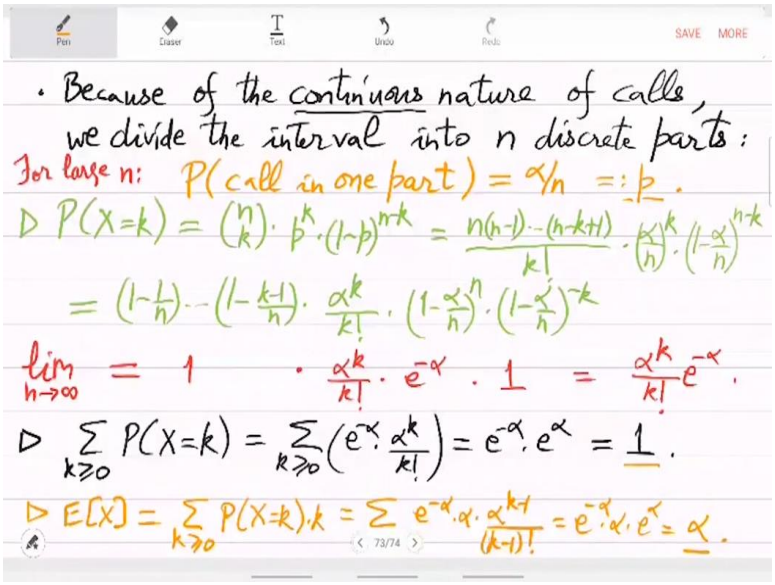
Probability for Computer Science
Prof. Nitin Saxena
Department of Computer Science and Engineering
Indian Institute of Technology - Kanpur

Module - 4
Lecture - 14
Concentration Inequalities. Variance.

(Refer Slide Time: 00:13)



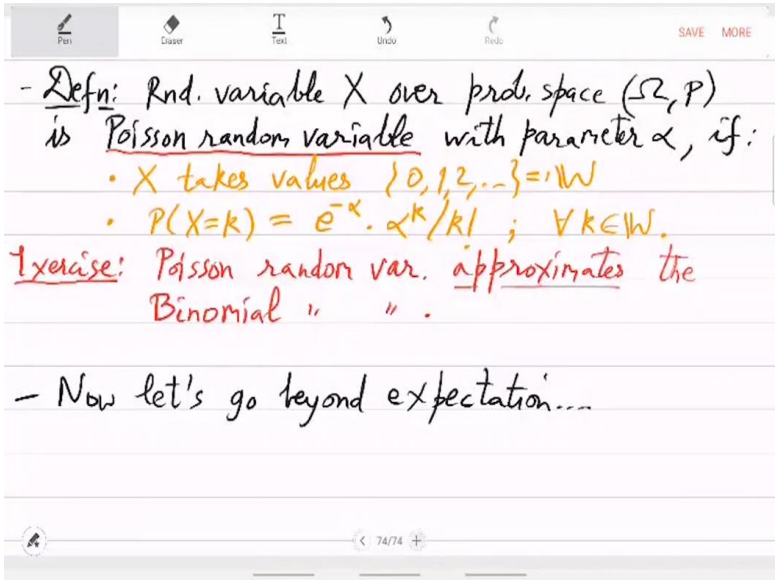
One thing you have to remember here is that, this alpha by n is probability, only when n is large. If suppose you take n to be alpha or even less than alpha, then this expression is more than 1. So, you cannot call it probability. So, here we are already assuming that this n is large. And then it makes sense that alpha by n is probability, when n is very large; ultimately, we make n to be infinity.

So, probability that X is some number, this actually comes out to be 1, which makes sense. So, this is a good consistency check. And just out of curiosity, what is the expectation of this random variable X? So, this is probability X equal to k times k, which is e raised to minus alpha times alpha to the k - 1 over k - 1 factorial. So, e raised to minus alpha times alpha, you can take out, and the rest is again of the form Sigma alpha to the k by k factorial.

So, it is again e raised to alpha. So, you get e raised to minus alpha; alpha out; and then, e raised to alpha, which is alpha. So, both the expressions are intuitively correct. Expectation

remains alpha and the sum over all possibilities remains 1. So, this is the perfect time to make this a random variable. This gives you the definition. So, let me write that down.

(Refer Slide Time: 02:06)



So, random variable X over probability space Omega, P; Omega, sample space; P is the probability distribution function. This will be called Poisson random variable with parameter alpha, if the following happens. So, first is that, takes values 0, 1, 2. So, basically whole numbers or non-negative integers. And second is that probability that it is k is e raised to minus alpha times alpha to the k by k factorial.

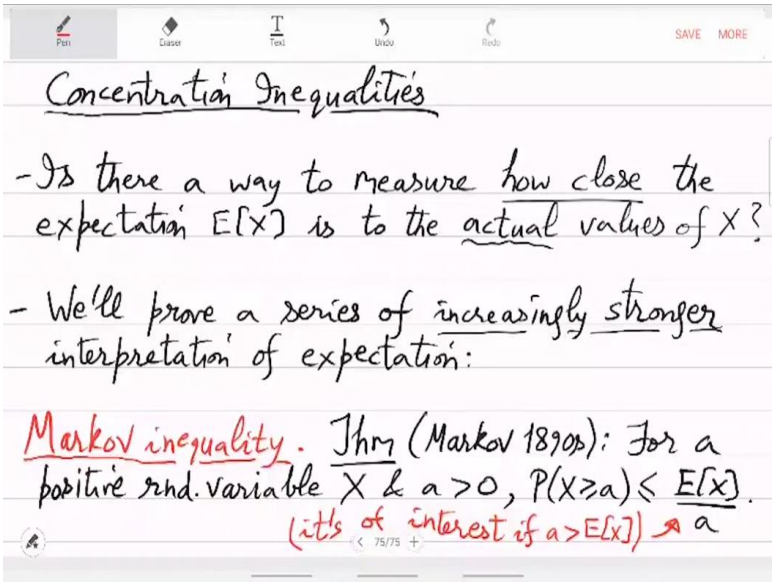
And you have already seen that this expectation of this is alpha. So, if you sum up over all the possibilities, it is expected to take value alpha; but a particular value k, it will take with this much probability. So, this is for k greater than equal to 0, in this domain. So, let me call this domain W, and that is where k is; whole numbers. So, that is your mass function; probability mass function is this.

I will leave it as an exercise, which you already saw actually; you saw the calculation that, this mass function is very much like the binomial mass function. In fact, you can show formally that Poisson random variable approximates the binomial one. Why is that intuitive? Because you saw the expression; so, you had this binomial random variable expression n choose k p raised to k 1 minus p raised to n - k.

And then you wrote p as alpha by n and you got this, for a large n. So, the similar thing; so, if p is very small, if p is significantly small compared to n, then you can do this calculation for

an appropriate alpha, and you will get; that Poisson is actually binomial in the limit. So, let us now shift gears; let us now do something else which is going beyond expectation. Now, let us go beyond expectation. So, that will be concentration inequalities.

(Refer Slide Time: 06:22)



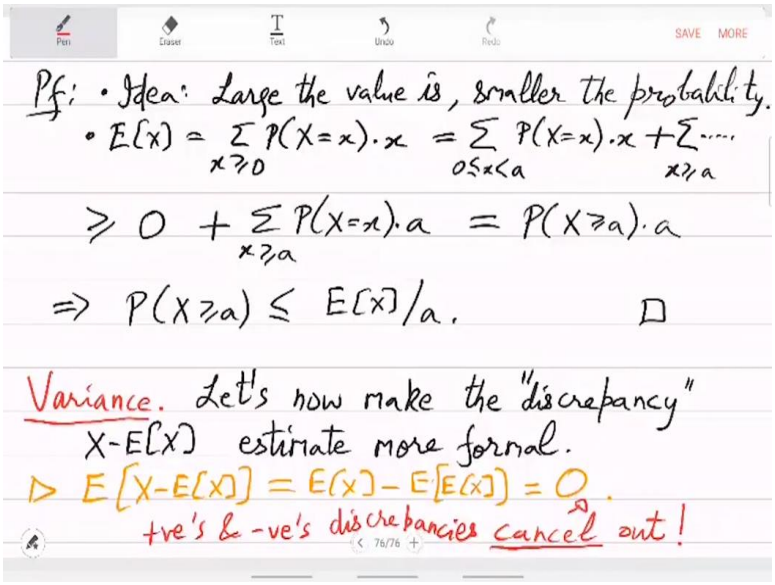
So, what are they? The question we want to answer in this topic is, is there a way to measure how close the expectation E X is to the actual values of X? So, what does it even mean? It means, questions of the type that, what is the probability that X is large, let us say, twice the expectation? So, what is the probability that X will overperform? Or, what is the probability that X will underperform? So, those tail probabilities, we want to calculate.

So, is this possible? Is expectation a good judge of that, or a good indicator towards that? So, in fact, it turns out there it is; and we will prove a series of increasingly stronger interpretation of expectation. The first thing we will show is called Markov inequality. So, the theorem is by Markov from 1890s. So, it says that, for a positive random variable X and a positive a, the probability that X is at least a.

So, think of a as being large. So, what is the chance that X is large? So, that surprisingly is related in a simple way with expectation. So, this probability is smaller than expectation by a. So, this is of interest only when; so, it is of interest if a is greater than the expectation; otherwise not. So, only when a is larger than expectation, this will give you a useful information; otherwise, this will itself be more; it will be saying that probability is less than equal to something that is bigger than 1, which is useless information.

But, if you take a to be, let us say, bigger than expectation, then this is saying that the probability of that happening is low. And it is lower as you increase a. So, that is a very interesting property. And how do you prove it? Actually, the proof is quite simple.

(Refer Slide Time: 11:22)



So, proof idea is just that; so, larger the a is or larger the value is, smaller its probability will be. How do you implement this idea? How do you see this in formulas? So, you write expectation by definition as big X taking value small x for all, in the non-negative real numbers. So, you break this up into below a and above a; a is positive, right. So, below a, this summoned; and then, the same summoned above a; small x at least a.

And look at the lower bound; so, when small x is less than a, in that case, you use 0; obviously, it is at least 0. And when small x is at least a, then you use a as the lower bound; so, that will give you probability of X small x times a, small x at least a. Which is what? Which is probability that big X is at least a. So, by definition of expectation, you already get this lower bound, which means that the probability that X is large is smaller than the expectation divided by a.

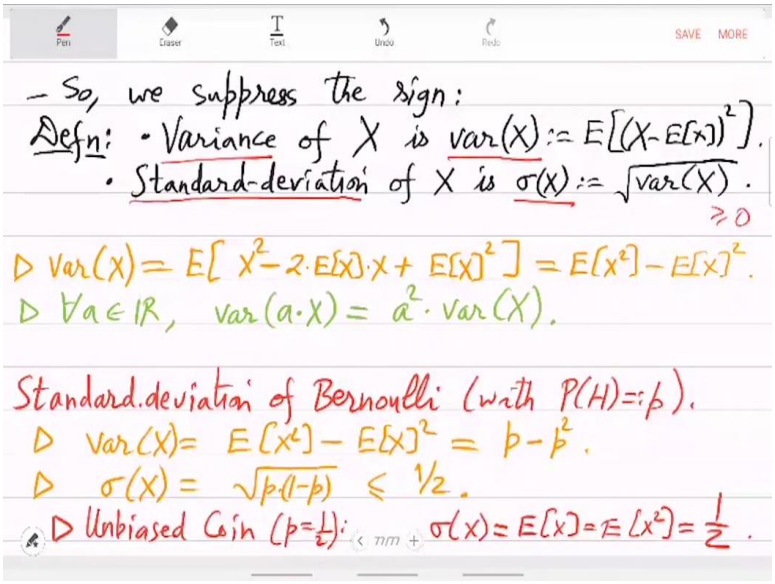
So, all these, the large part, the summons where small x is large, they, since a is large, the probability part has to be smaller. So, this exactly is the interpretation in the definition of expectation; but for this, you need a large a; small a, this calculation is useless. And now, going beyond expectation, that function is called variance. So, let us now make the discrepancy.

Discrepancy of X, we want to understand it from expectation more carefully, more precisely. This previous inequality was of limited use. For example, it does not tell you the probability that big X is half of expectation. So, we want to ultimately reach that stage. And you want to be more precise. So, what is the difference between X and expectation? So, let us make that discrepancy, which is X minus E X, this discrepancy measurement or estimate more formal.

So, the way to do it is, just look at the expectation of this discrepancy. What happens? So, if you look at the discrepancy expectation, you get into a trivial information by linearity of expectation; because, this is nothing but expectation of X minus expectation of X. But the latter thing is a constant; so, it is just expectation minus expectation; so, it is 0. The reason why we got nothing is because the positives and negatives cancel each other out.

That is what happened. Sine was the problem. So, you could not learn anything about the magnitude of the discrepancy. So, one solution that people found long back is to square it and then calculate the expectation.

(Refer Slide Time: 16:54)



So, we suppress the sign by using this definition. So, variance of X is expectation of X minus; so, discrepancy squared. Let us work with this. It is not directly 0; it is not clear what this value is. It will depend on the random variable. And since you are actually interested in the discrepancy and not its square, we can take the square root of this; and that is called standard deviation.

So, we will call it sigma X. Compute the variance; take the square root, positive square root. So, that is the positive square root. Notice that variance will not be negative, it is non-negative. So, square root will also be non-negative; it is a real number. So, this seems to be the first right notion to understand or to measure or to estimate the discrepancy; how far is the random variable from its expected value. Let us do some calculation.

So, variance using linearity of expectation, what you can do is, you can expand out. So, it is expectation of X square minus 2 E X X plus E X square, which is; so, by linearity, you will get expectation of X square minus twice expectation of X times expectation of X. So, that is expectation of X whole square. So, it is a kind of the average of square minus square of the average. That is variance.

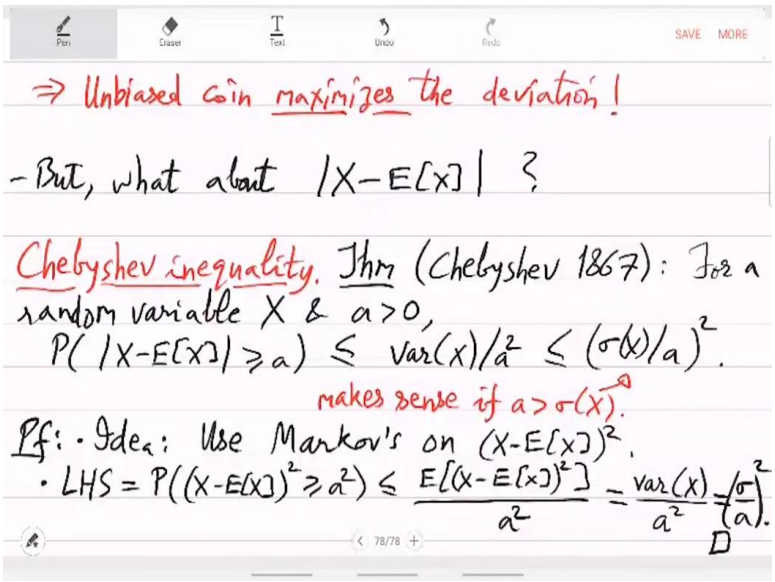
And one more thing you can prove quite easily; for a constant; then you look at the random variable multiplied by a, that constant. That, you can see from this expression or even the definition that an a square will come out; that is again by linearity, by scaling of expectation. So, you will get a square times variance of X. So, those are the properties that you immediately learn from the definition of variance. Let us see an example.

So, standard deviation of Bernoulli random variable; what is a Bernoulli random variable? So, recall that it is a, just a coin toss, single toss, with probability of heads being p, parameter p. That is quite simple. So, variance of X is expectation of X square minus expectation of X whole square. Now, this Bernoulli variable X, that takes 0 or 1 with probability p and the 0 with 1 - p; so, expectation is p. So, you get p minus p square.

And the standard deviation you get square root of p 1 - p; which you can also show is at most half. That is, you cannot increase this standard deviation beyond half. So, what you can prove now, in the case of an unbiased coin, that is p equal to half, you can show that standard deviation expectation of X and expectation of X square, all these values are half. For p equal to half, you get this.

And so, for an unbiased coin, the standard deviation of Bernoulli is maximised, which makes sense because the coin is unbiased. So, that is where there should be no bias either side of the expectation which is half. So, it wildly oscillates from 0 to 1, discrete; either X is 0 or X is 1 with equal chance.

(Refer Slide Time: 23:25)



So, unbiased coin maximises the deviation. So, that is good. We have some understanding of discrepancy and deviation. But, could we go back to the discrepancy X - E X magnitude? What can we say about that? Not square of this; square of this, we understand better, that is variance. What about this, just the magnitude? So, that is covered under now Chebyshev inequality. This is by Chebyshev; was the advisor of Markov actually, from 1867.

So, for a random variable X and positive a, he showed that the probability of this discrepancy being large is small. How small? Depends on the variance; variance by a square. Variance is a standard deviation square; so, you can write it as sigma X by a whole square. And again, this will make sense when a is bigger than the standard deviation. Otherwise, it is a trivial claim; it is simply saying that the probability is less than equal to 1 or more, which everybody knows, which is clear from the definition.

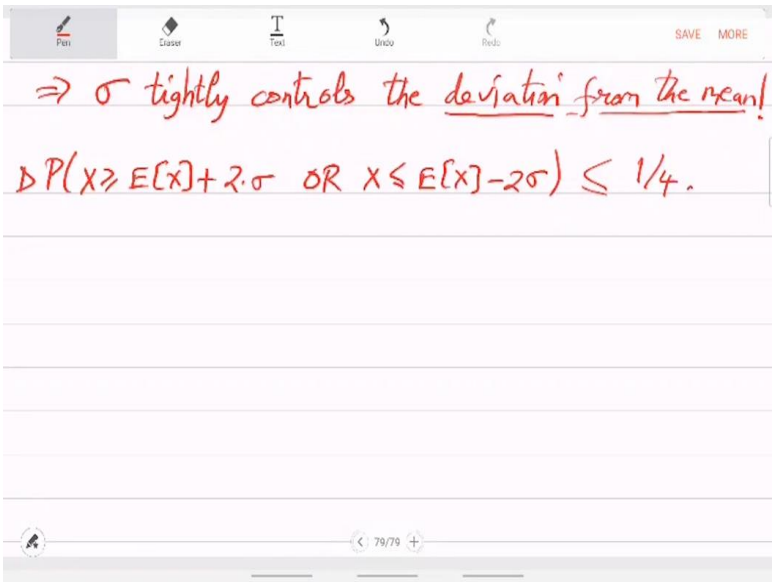
But if a is, let us say, double the standard deviation, then this is saying that the probability is less than equal to one-fourth, which is very surprising, because you actually are getting a quantitative value. It is a quantitative fact about this intuitive thing that X cannot be very far away from the average. So, this is very precisely telling you that X cannot be, the discrepancy cannot be more than double that of standard deviation. The probability is low.

So, this proof is also now very easy, given your background. So, the idea is, use Markov's inequality on variance or on this square of this discrepancy. So, how will it go? So, the left-hand side which is the probability is equal to the same, to the probability of X minus

expectation square being at least a square. Why is that? Because a is positive; and left-hand, this discrepancy is also positive. So, you can as well square it; it works.

So, probability of this event is equal to the probability of original event, which by Markov is now expectation of; square of this divided by a square, which is nothing but variance by a square, which is also standard deviation by a whole square. So, that is what we wanted to show.

(Refer Slide Time: 28:37)



So, what this tells you is, this sigma, it tightly controls the deviation from the mean. So, the random variable, when you do an experiment, it cannot be very far away from the mean. That is what deviation is controlling. So, for example, it is saying that, probability that X is away from the mean by 2 sigma or probability that X is smaller from the expectation by 2 sigma, either of these events combined, their probability is smaller than one-fourth.

So, the probability is less than 25%. That, you do an experiment and you get 2 sigma over the mean or double the standard deviation below the mean, that is a very small probability. These are low probability events.