Algorithms for Big Data

Prof. John Ebenezer Augustine
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 11
Tail Bounds

What we are going to do here. So, usually tutorial you know all some people here might know more than others (Refer Time: 00:32) to present it in a way that will be of interest all of you, whether you know able to what are talking about; whether it is all completely new for you, and since I do not have a clue what you know I really encourage you to ask questions. And this is what I tell my students if you teach you do not understand something and you are shy, hesitate to ask because you show that everyone else actually know they do not. So, someone has to ask.

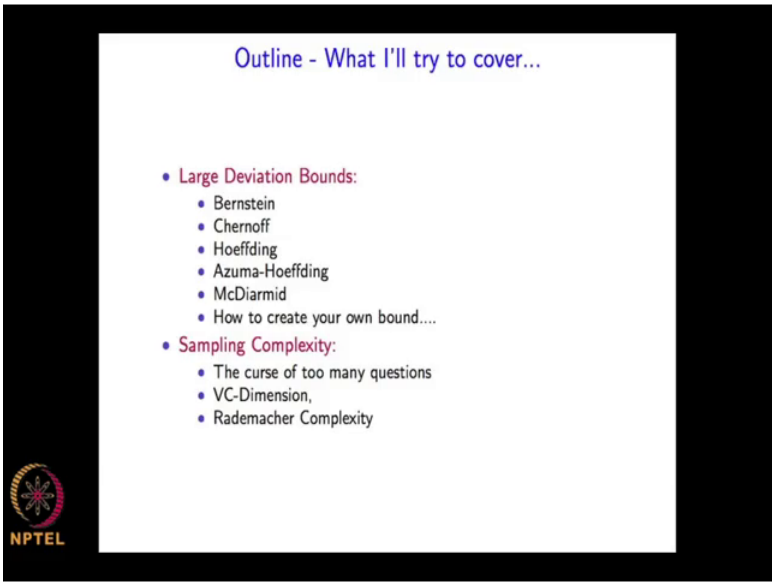
So, what we are going to talk about is, a kind of the very basic question of probability adopted to the particular application that we need in computer science or we get in few slides to the question of why we develop in the (Refer Time: 00:35) kind of new probability tools in computer science will probability whose all with therefore, many years and I will try to answer that question. But basically what we are going to look around is we go to start with a (Refer Time: 01:52) large deviation bound.

Now, assume that some of you know, have used over this trade off something called the Chernoff Bound. So, if you have never heard about the Chernoff Bound (Refer Time: 02:08) this is probably the most often used tool in analysis for algorithms these day, but if you know what Chernoff Bound, well I want you show you is that is I am going to give you some kind of perspectives or what is Chernoff Bound? So, most of us used Chernoff Bound is kind of black box well, if I fit all the requirement then I can use it in otherwise no I came. Well, that is not the real thing; Chernoff Bound is actually an example of kind of general scheme called large deviation bound.

Chernoff what we gave a popular in computer science, but they have the many importants violence which also used these days, but unless going let us known, and just for a fun of with I list of this as the donations list of the names of all the bounds. And there is along there is a long history here. But I Chernoff did not invent the Chernoff Bound, again often in science the names are note the right name and even chernoff himself someone you can still meet in bus stand his retired.

How a professor knows the computer science, but where is (Refer Time: 03:21) he would tell you that it is all the mistake he never did it, it is not credit to him. It is probability many (Refer Time: 03:28) time who is break in (Refer Time: 03:30) for me he was the first one to think about this idea, and then there is a whole chain of developments which will go one step after the other which goes Chernoff and Hoeffding, Azuma-Hoeffding, McDiarmid and they all have some meaning here.

(Refer Slide Time: 03:49)



Outline - What I'll try to cover...

- Large Deviation Bounds:
  - Bernstein
  - Chernoff
  - Hoeffding
  - Azuma-Hoeffding
  - McDiarmid
  - How to create your own bound....
- Sampling Complexity:
  - The curse of too many questions
  - VC-Dimension,
  - Rademacher Complexity

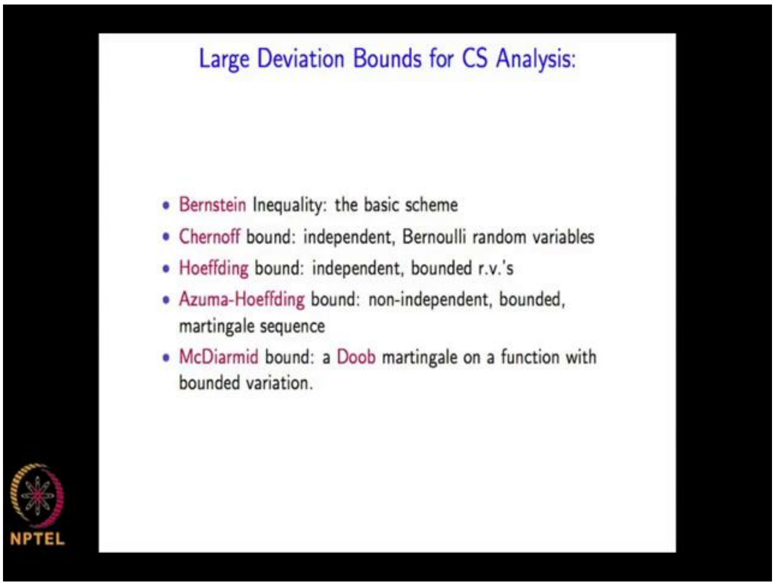
Then I will show you that they know this actually general recipe here, and if you understand the recipe then you can use it for any other application that do no t fit in to this, so that particular one.

So, large deviation is about how far is then empirical expectation, The empirical average from the actually expectation of random variable, which is they basic question you ask in statistics thing. So, I take a few samples. So, I want to know what is the average height of his student in IIT is. So, I measure random sample, and I take the average and now I ask. So, this is my static or this is the empirical number that I have for my measurement, but there is some real value the average, and the question is how far they are form each other. That is the basic question in statistics, and large deviation tries to answer it in waves question. Waves, but as we focus now in computer science any particular or the big success or one of the biggest successes of computer science in this India is machine learning.

And in machine learning, the question is much more complicated than how far is my sample from my real value, because I often do not will you know what I am why searching for. In fact, that I want to sample well we go for many questions in simultaneously. So, these (Refer Time: 05:34) on the again many names, will refer to it particular computer science refer to as the sample complexity, it also something called uniform convergence, many names for these and it has (Refer Time: 05:49) students well, but became very important in recent computer science. So, the second part of what are present here will talk about this sampling complexity we will be today we will take a few hours before we get them.

That is kind of high level tools about what are like to cover here, and yeah to make some advertisement you know it is a very high (Refer Time: 06:14) you have to commercials. So, yeah it is all part of it, it is in a book which some of you might know and some of with is a (Refer Time: 06:23) and it will appear in the book, that is why can I finish the work to write in a second edition, but as you know it takes a years until that is you can see it in book. So, it is all there in a more will be available soon.

(Refer Slide Time: 06:07)



Large Deviation Bounds for CS Analysis:

- Bernstein Inequality: the basic scheme
- Chernoff bound: independent, Bernoulli random variables
- Hoeffding bound: independent, bounded r.v.'s
- Azuma-Hoeffding bound: non-independent, bounded, martingale sequence
- McDiarmid bound: a Doob martingale on a function with bounded variation.

So, so large deviation bound is kind of trade off between and again it is a usual trade off you have been sign in mathematics, were if you give me more conditions in the theorem, I can proves stronger results. So, a lot of theorems about the trade off of between what you have to assume and what can you proof with this assumptions. So, the same game is in the Chernoff Bound in the last deviation in particular of the the most popular of the Chernoff Bound basically talks about independents the bernoulli random variable independence 0, 1 random variable.

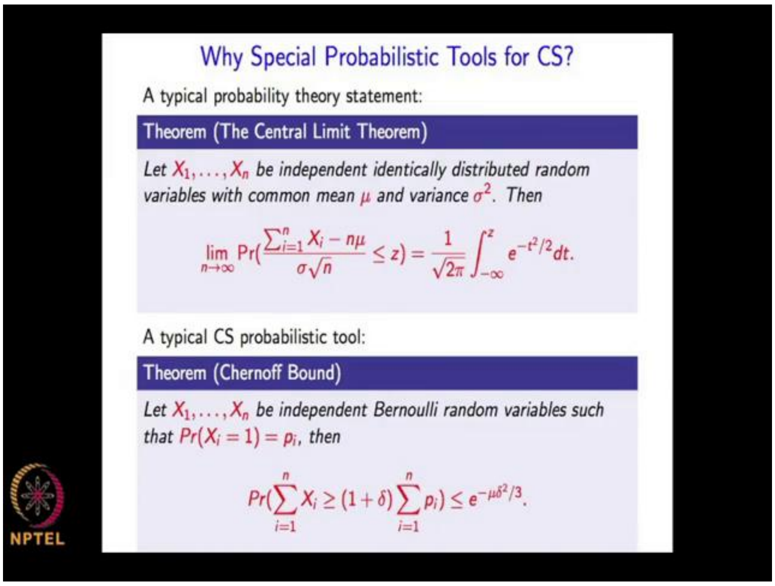
So, if I take (Refer Time: 07:37) ask how many independence 0, 1 random variables I have to average or values I have to average in order to get a good estimate of the actional probabilities, which is 0, 1 random variable the expectation is the probability. Now you say well ok, life is not only 0, 1 and the random variable they are more interesting things.

So, the first kind of relaxation, who says well; so what are I have independent observation of a random variable, but it is not 0, 1 random variable. Well it is not have to show that if it is completely unbounded this nothing you will do, but if the random variable is bounded, then will get a bounds which handle that one and this is called Hoeffding bound, and then come believe jump; what you say well do they really need to be independent? So, well we will actually get rid of that is well, while in some way so, what we will do is we will talk in I am sure that this will be new to some of few and not new some other people, but I will talk a little about something called martingale, which is sequence of random variables that are not independent, but they have very natural dependence and we will find the way to found to prove a variation bound though a last deviation bound even for a sequence of random variables that are not independent.

And then we will see that from that one we get very useful results about bounding to deviation random of a function of many random variables, and then suddenly it will jump to questions like what is the chromatic number of a random graph or what is the links of the past.

So, suddenly you will get from probability to actually answering concrete question you know seems completely from a different world. Where we get before we get the others these question that always ask when, I give this kind of tutorials in often a when the audience is may be not like here, but audience thus varies from different disciplines and the obvious question is was well you are computer scientist and you think you event at the world, but you know the well interesting science before computer science. So, mathematician, probabilistic, researchers already ask all these questions. So, why we invent we will why we build new theory of probability, or probabilistic technique for computer science. And you could have a long answer for that, but the short answer is here.

(Refer Slide Time: 10:26)



Why Special Probabilistic Tools for CS?

A typical probability theory statement:

Theorem (The Central Limit Theorem)

Let $X_1, \ldots, X_n$ be independent identically distributed random variables with common mean μ and variance σ^2. Then

$$\lim_{n\to\infty} Pr(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt.$$

A typical CS probabilistic tool:

Theorem (Chernoff Bound)

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables such that $Pr(X_i = 1) = p_i$, then

$$Pr(\sum_{i=1}^{n} X_i \geq (1+\delta) \sum_{i=1}^{n} p_i) \leq e^{-\mu\delta^2/3}.$$

So, here is there prototype of result in statistics or in probability theory. So, probably recognize The Central Limit Theorem, and the central limit theorem basically (Refer Time: 10:42) n independent, identically distributed random variable with mean mu and variance sigma square, then if I take a lot of observations. So, if I take the average of this, and normalize it in a while then what do I show? I show that in the limit this distribution tends to be normal, often you also learn. Well that is way is nice for probability mathematics, where in computer science the limit is not enough, I want to know How far I have go.

So, in computer science we really want to counts, we want to count the steps. So, each of this sample is as a cost. So, in (Refer Time: 11:37) it as a cost I have to go and you know sample ask someone a question or measure or something. In computer science that is computational cost.

So, it is (Refer Time: 11:48) to know that in the limits, you know that the variation you know behave like normal distribution, where there is an help me I many life when you ask me how many samples are you need. So, in computer science we want to have something that we do not care about the limits. We care about you know half we have to go, you know to be sufficiently close to the limit. In other words this is these Chernoff

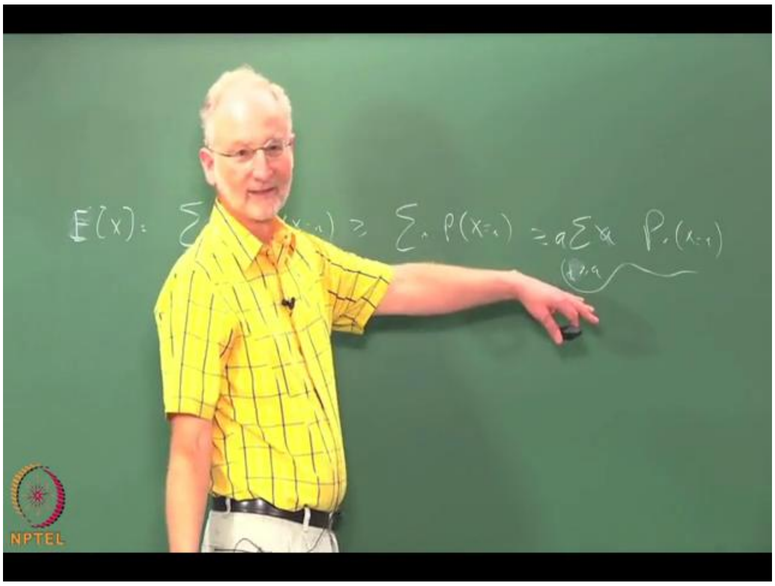
Bound will get to this in detail, but just to give in flavor of what we going to do here and they contrast to the central limit theorem.

Here we are going to say if I have n independent Bernoulli random variable observation, observation of Bernoulli random variables. Well if there are (Refer Time: 12:38) distributed in the probability of which of them being one is pi i. Then the probability that is they some of them, is pi from the sum of the expectation, is going down according to the expectation of the sum. If we may be better to normalize is to say why the probability that the average here, is (Refer Time: 13:17) by delta from the expectation of the random variable and then will get those exponentially down exponential use n.

So; in other words, in not only how where is the limits, but how far do I get to the limit. So, that is the big difference between many of the tools we made in computer science, and the tools that you use in that you get in standard probability theorem. (Refer Time: 13:47) it is also well we do not talk about here, but also (Refer Time: 13:49) we Queueing theorem. So, in Queueing theorem you have the beautiful results, but usually talk about the limits distribution, and then when in computer science we want to ask about the size of buffers, the queues, in routing and stuff without, then Queueing theorem has somewhat limited in answering a question. Because we want to know you know how first things converge what happen when they do not converge and stuff like that once a side remark.

So, now we are going to ask. So, now, you have to stop. So, this was introduction now you get to the stuff. Now everyone in probability is some for Asian on what is called Markov Inequality. Now that is the very trusting phenomena, Markov Inequality, on it is own is extremely weak bound, which you probably also in some class it is extremely weak bound, but everything is building on it. So, what is a Markov Inequality it only applies to random variable that is non negative, and the probability that is random variable is greater than; so value a, is bounded by the expectation of the random variable divided by this value, a and it is really (Refer Time: 15:19) want to proof if you think about the expectation, I am sorry.

(Refer Slide Time: 15:26)

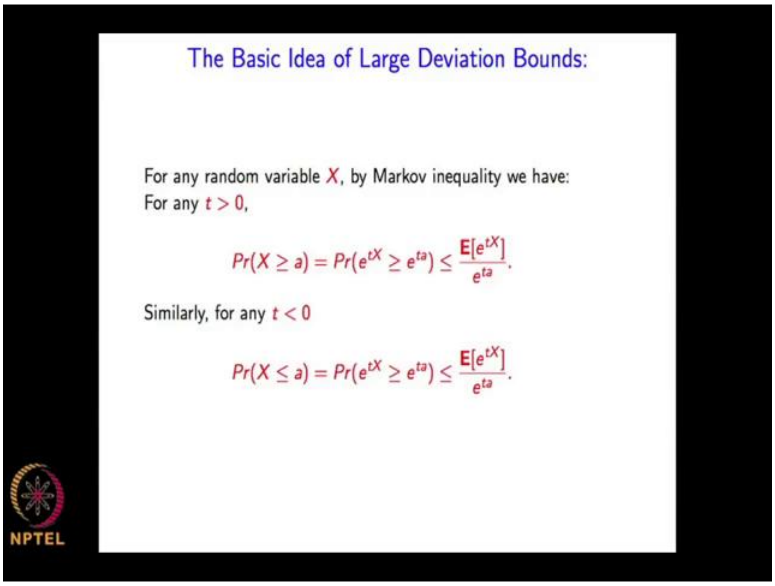


The expectation of x, let us assume that is discrete and the variable. So, that is sum of I, I probability that x equal i. So; now, let us do two simplifications. First will just look, this is definitely greater or equal, than if we just sum above a probability that x soon time I probability that x equal i, but now we sum only above a then, definitely is greater than sum of a probability, I greater or equal to a probability x equal i.

So, now if we take a outside, then what is we can here the probability that x is greater than a, and we get this. So, that is a very simple bound and only we the only reason we mention it here is, because everything else will do, will built on that bound. So, amazing that you start with this will show later will do the comparison with other one see that this is so weak, but when you built on this you get amazing (Refer Time: 16:55) stuff the first one again note really what we will talk about here, where the first bound that is built on Markov is Chebyshev’s Inequality and Chebyshev’s Inequality basic said well instead of thinking about x now thinking about x minus e of x square and then we plug it in and we get Chebyshev.

Chebyshev is (Refer Time: 17:19) for a stronger bound, it works for random variable that oppositive and that can be positive or negative and again will do the comparison we will see nothing it get better one. Perhaps note do not get (Refer Time: 17:30) to be needed.

(Refer Slide Time: 17:35)



The Basic Idea of Large Deviation Bounds:

For any random variable X, by Markov inequality we have:
For any $t > 0$,

$$Pr(X \geq a) = Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

Similarly, for any $t < 0$

$$Pr(X \leq a) = Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

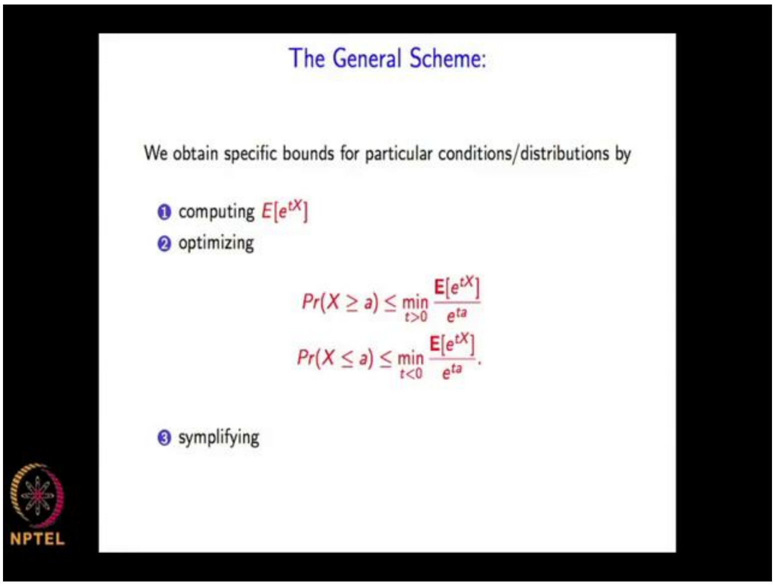
We want bounds that go down exponential. Now, we get you this general technique of large deviation bound. The technique is the following; I want to know what is the probability that x is greater or equal to a, well if I pick t that is greater than 0, then the prob h is equal to a is the same as the probability that e to the tx is greater than e to the ti. As long as I took t that is greater than 0; now we got something about interesting, we started using random variable that can be positive or negative, but e to that is is always positive or non negative.

So, now I can take this, now we run the random variable here that is no negative and now I can apply Markov inequality, but Markov inequality now says what is the expectation of this divided by this. So, the probability that this random variable is greater than this is bounded by the expectation of this random variable. So, we took random variable, and instead of asking what probability these are the variable is greater then I, we ask what is probability that e to the t to the x e to the tx. So, of actually this random value is greater than e to the tx.

Now, if you look at it that is that is a quality. So, we can of did not do anything rather than retain it in a more complex way. As it happened this jump for looking at the random variable looking at the exponent of the random variable as a very dic meaning, that we have to discuss in second. This gives very strong bounds. Now this is for the question whether x is greater or equal to a. e can do we can ask. So, the opposite question is using t less than 0. So, the probability that x is bounded by a from below a from above well now t is negative.

So, if multiply t both sides we switch the inequality. So, this is now going to be looking the same, only t is negative and we get the same bound using the Markov Inequality. Now realize for those of you see it the first time this is just symbolic manipulation and you will take some time and until I can show you, know they application of this. So, there was (Refer Time: 20:41) we have to go some formalities, until we get to the real application, but this is the kind of the general form.

(Refer Slide Time: 20:51)



The General Scheme:

We obtain specific bounds for particular conditions/distributions by

1. computing $E[e^{tX}]$
2. optimizing

$$Pr(X \geq a) \leq \min_{t>0} \frac{E[e^{tX}]}{e^{ta}}$$
$$Pr(X \leq a) \leq \min_{t<0} \frac{E[e^{tX}]}{e^{ta}}.$$

3. symplifying

So, in order to get all these different bounds that we look before that we enumerate before, basically what you doing is you doing the following. So, here is like a rest of p, an algorithm for generating bounds. So, you take random variable, they have part usually is to compute the expectation of e to the tx. Now you gets for what we had before that prob at x is greater than a, is bounded by this expression and the probability x is bounded below by is this expression.

But now; where we free value here to play with, because here we just say well take any t either greater than 0 or lower than 0 , smaller than 0 . Now, you can play with t, so of course, will go will try to find the t, since we want to get a bound we try to play was t to find the minimum value for this bound. Now, we the next second step, that is often give us kind of how to remember bounds. So, this third one is usually we simplifying to something that is easy to work with. So, all the bounds will have this structure.

But before we get to particular bounds, but I want you to show you and this is often it grow completely in computer science, is now this is not the magic there is very particular reason, why we move from the random variable to the exponent to the random variable in the exponents, so where we those of few who study enough probability, what is this function.

Student: (Refer Time: 23:03)

Hm?

Student: (Refer Time: 23:06).

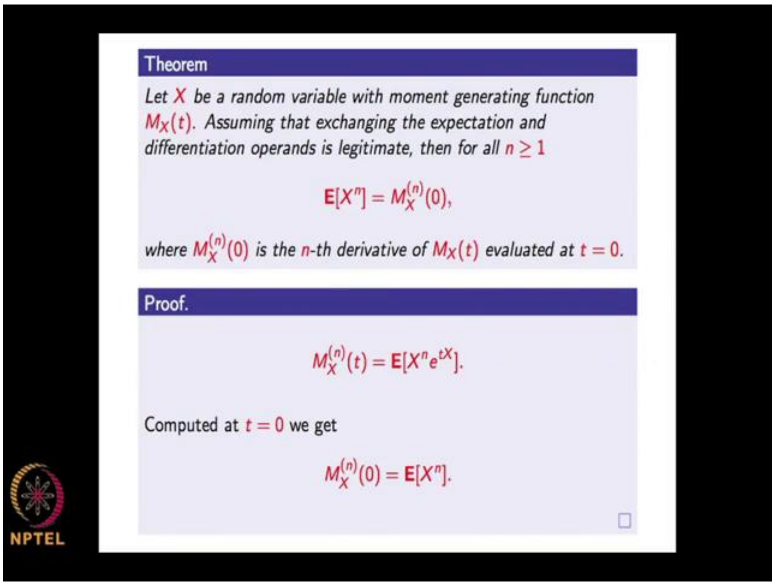
I do not hear.

Student: Moment generating function.

Moment generating function, it is one of that is Chernoff of being (Refer Time: 23:16) we there may about some function who tries a long list of function; there is a particular reason why we use the exponents of the x. The reason is that this is the moment generating function, and the moment generating function in a way and called the whole distribution.

So, Markov Inequality is to say well give me the expectation and I will give you some bound. Chebyshev said well give me a little bit more, give me the variance and I will give you stronger bound. Now we say, I do not want to just to first moment the second moment give me all the moments. Give me the all distribution. Well, if you give me the all distribution, I give you much, much stronger bounds. And I will show you in a second how strong they are.

(Refer Slide Time: 24:16)



Theorem

Let X be a random variable with moment generating function $M_X(t)$. Assuming that exchanging the expectation and differentiation operands is legitimate, then for all $n \geq 1$

$$\mathbf{E}[X^n] = M_X^{(n)}(0),$$

where $M_X^{(n)}(0)$ is the n-th derivative of $M_X(t)$ evaluated at $t = 0$.

Proof.

$$M_X^{(n)}(t) = \mathbf{E}[X^n e^{tX}].$$

Computed at $t = 0$ we get

$$M_X^{(n)}(0) = \mathbf{E}[X^n].$$

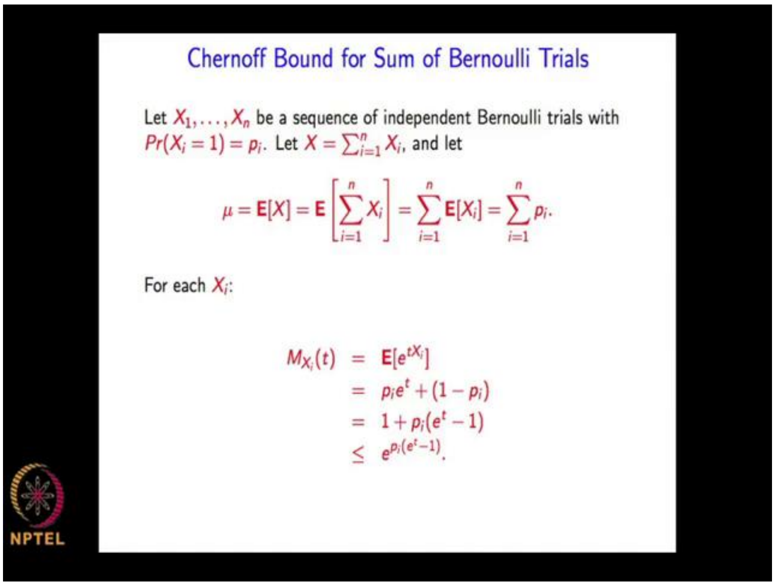
□

So, just to remind you; so the moment generating function is the expectation of e to the tx, and this is just (Refer Time: 24:23) remember. So, the moment generating function there is we call the moment generating function is that, in the function encoded in the function. All the moments of the random variable, moment of the random variable is the expectation. So, a moments, so they the k moment of a random variable is E of x to the k and a really how theorem to proof that usually do not do in a basic causes is that.

There is one to one correspondence between a distribution, and it is sequence of moments. So, if you know the sequence, well if all the moment exists. Then knowing the sequence of moment is a fluent to knowing the distribution, you can go for one to the other. So, the all generating function defines all the moments. So, in a way for the distribution that we work is for which all the moment exist, in a way what you put into the recipe what you put into the inequality is the food information about the function, but the distribution.

Or you do it well if you take the moment generating function, you take the k the revertive of it is and you sets t to 0 then you get the t moment the k moment. They last thing that the I want to you remind you is that is the moment generating function of a sum of random variables, is the product of the moments for of independent, the sum of independent random variable is the product of the moment generating function. So, that is actually very simple because if you write it is. So, the moment of x plus y is e to the t, x plus y and if it is independent. Then the expectation you can take the product of the expectation in the term, will just going to use it in second.

(Refer Slide Time: 26:44)



Chernoff Bound for Sum of Bernoulli Trials

Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli trials with $Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$, and let

$$\mu = \mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum_{i=1}^{n} p_i.$$

For each X_i:

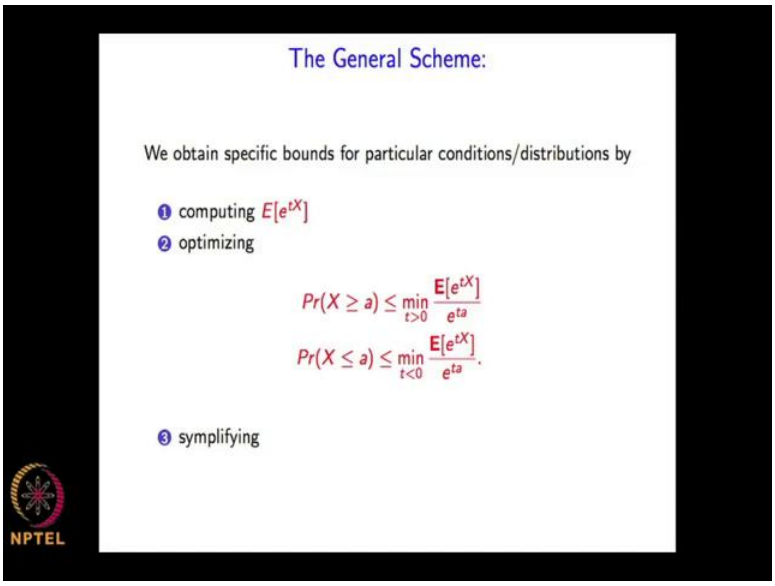
$$\begin{aligned} M_{X_i}(t) &= \mathbf{E}[e^{tX_i}] \\ &= p_i e^t + (1 - p_i) \\ &= 1 + p_i(e^t - 1) \\ &\leq e^{p_i(e^t - 1)}. \end{aligned}$$

So, one thing to remember is that is it is not the magic that we go to this e to t x, we using the moment generating function in our expect to get better stronger bound.

So, it is a kind of the proof of the first basic Chernoff bound. So, let us x one to x n be a sequence of independent Bernoulli prior random variable. So, the probability x is 1 is p I probability x I is one is pi the probability, and then we looking at sum of xi's. So, notice (Refer Time: 27:27) way that is we no need the x 1 to x n to be identically distributed what we need them to be 0, 1 random variables, but each of them can have it is probability to be one. Which and (Refer Time: 27:42) you also these expectation. I was have a some this is actually very important, when we get to analyze algorithms because we often ask well how many steps we have to do it is steps some probability of success or not. So, if you know we can still apply the Chernoff Bound even if the probabilities of defect.

Now let us, mu be the expectation of the sum. Sun of the expectation which is the sum of the pi's. Now, both we want to ask is if I take n observation of x 1 to x n, and I look at the sum the value of what I get, how far is it is from sigma of pi. In other words if I look at the sum of an observation, how good are the as a predictor for the read expectation that is shuffling it is statically question.

(Refer Slide Time: 28:56)



The General Scheme:

We obtain specific bounds for particular conditions/distributions by

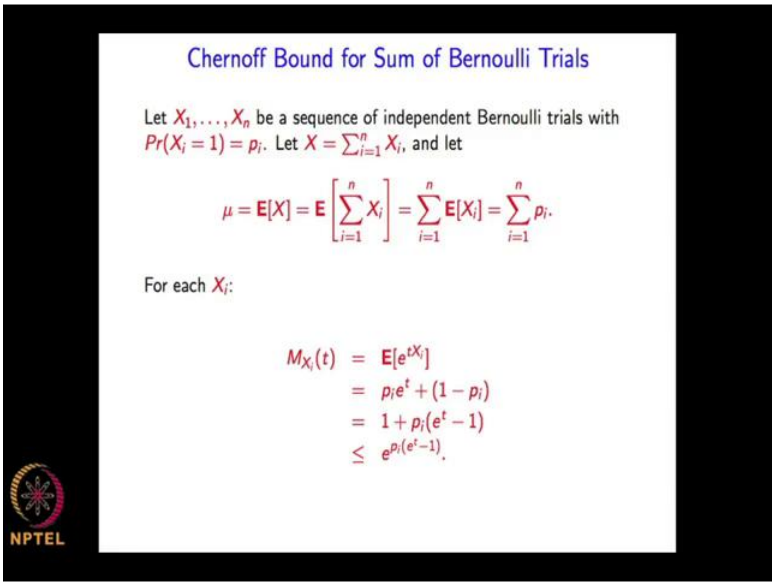
1. computing $E[e^{tX}]$
2. optimizing

$$Pr(X \geq a) \leq \min_{t>0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}$$
$$Pr(X \leq a) \leq \min_{t<0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

3. symplifying

So, is that we use this recipe and the first one is we have to computes expect the moment generative function or the expectation of e to the tx, that is usually whether difficulties.

(Refer Slide Time: 29:14)



Chernoff Bound for Sum of Bernoulli Trials

Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli trials with $Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$, and let

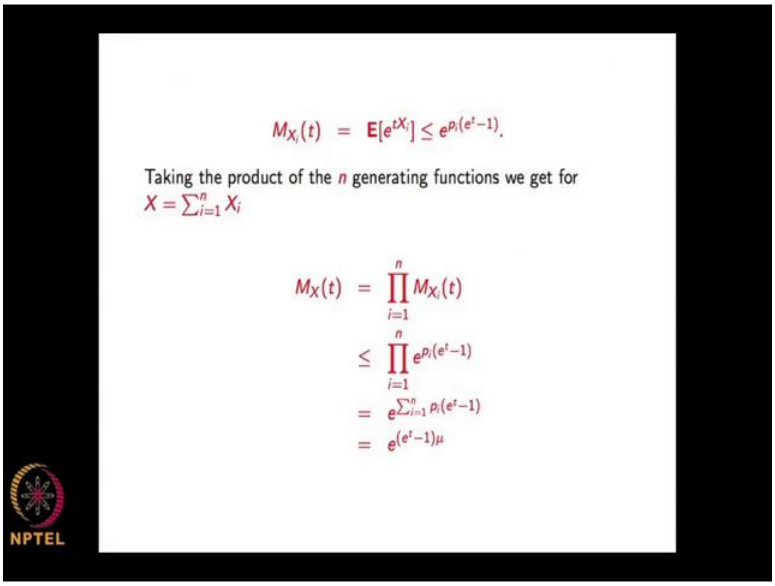
$$\mu = \mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum_{i=1}^{n} p_i.$$

For each X_i:

$$\begin{aligned} M_{X_i}(t) &= \mathbf{E}[e^{tX_i}] \\ &= p_i e^t + (1 - p_i) \\ &= 1 + p_i(e^t - 1) \\ &\leq e^{p_i(e^t-1)}. \end{aligned}$$

But now, we going to use the fact that x is sum of independent random variables. Here some of independent random variables. So, in that case we just have to compute the expectation of each xi of e to the txi or, but now it is getting very easy because xi just getting the value 0, 1. So, if it is one then it is pi e to the t and if it 0 it just one minus pi; that is easy and now we do some manipulation here and it is bounded by e to the pi, e to the t minus one.

(Refer Slide Time: 30:01)



$$M_{X_i}(t) = \mathbf{E}[e^{tX_i}] \le e^{p_i(e^t-1)}.$$

Taking the product of the n generating functions we get for $X = \sum_{i=1}^{n} X_i$

$$\begin{aligned} M_X(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &\le \prod_{i=1}^{n} e^{p_i(e^t-1)} \\ &= e^{\sum_{i=1}^{n} p_i(e^t-1)} \\ &= e^{(e^t-1)\mu} \end{aligned}$$

So, now, I have a bound for each of the moment generating function each of the xi's. So, I think the product to it is, and I get a bound for the sum and throughout this stock I will skip the arithmetic’s; so whatever the symbolic manipulation is all on the slides, if not in the book. I want to you give you high level idea I mean I do not want to get in to the (Refer Time: 30:32) computation. So, after I computed the moment generating function of each xi, I used fact that the x I are independence. So, the moment generating function of the sum is the product of the moment generating function. So, I finish one part.

Student: What mean mu here?

Sorry?

Student: Mu.

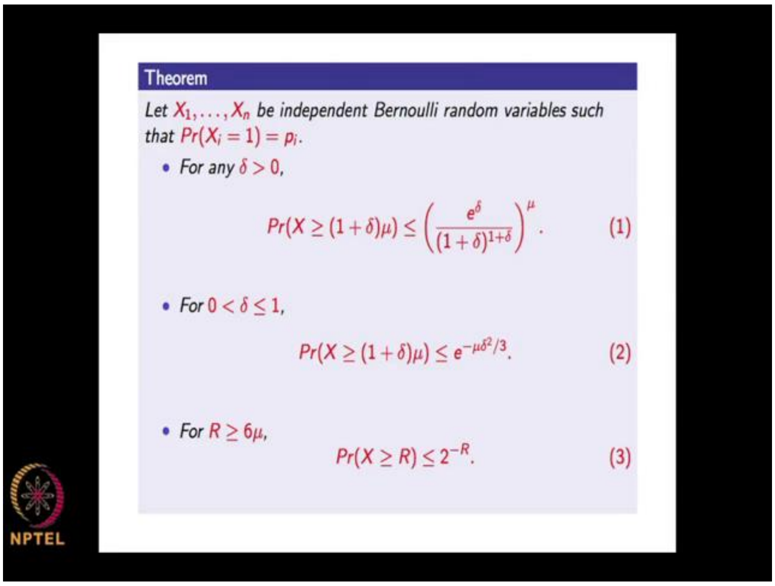
Use the expectation sigma of pi.

Student: So (Refer Time: 31:00).

So; now, we apply the Markov Inequality. So, the probability x is delta way from it is expectation. So, it is probability e to the t, tt x is greater than e t one plus delta mu, and now it is bounded by the expectation of this random variable divided by the value, and the expectation we computed is this value, and so, we have this value. Now, we said last part and second part of the recipe is, where we have this free value here t, free variable t and now you want to optimize with respect to t.

We want to get this value as small as possible with respect to t. So, you can take the derivative or just believe me that if you plug t to be lon of one plus delta you do some manipulation you get this bound. I have to see that is t only can plug you t that greater the 0, but if delta the deviation is greater than 0 then, t of lon 1 plus delta is greater than 0 and now everything is fine. I get first version of Chernoff Bound, it looks good, but it is somewhat how to remember or to memorize.

(Refer Slide Time: 32:51)



Theorem

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables such that $Pr(X_i = 1) = p_i$.

- For any $\delta > 0$,
$$Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}. \quad (1)$$
- For $0 < \delta \leq 1$,
$$Pr(X \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/3}. \quad (2)$$
- For $R \geq 6\mu$,
$$Pr(X \geq R) \leq 2^{-R}. \quad (3)$$

So, that come kind of games of are you take it is look just something that is easier to work is results, we know giving up too much on a bound.

So, this is title this is easier to remember. So, the probability this is what usually seen in papers, to the probability that is x deviate by more than delta from which expectation, is bounded by the expectation delta square over 3. Why you get this 3? So, just technicalities of how you show that these is bounded that these is an upper bound for this. Then you can get even weaker bond, but easier to work with that you go for a very large variation.

So, if you go for variation which is, 5 time the expectation then the probability that you greater than that value is that two to that to the minus that value, but you prove it you do arithmetically as basically high school calculus. So, you take a derivative the second derivative then you show that you get the right box. So, we will do it here. So, basically you show. So, for example, you want to show that these are always greater than this 1. So, take the derivatives in the second derivative you use some delta work and you get the bound, the same for the other one.

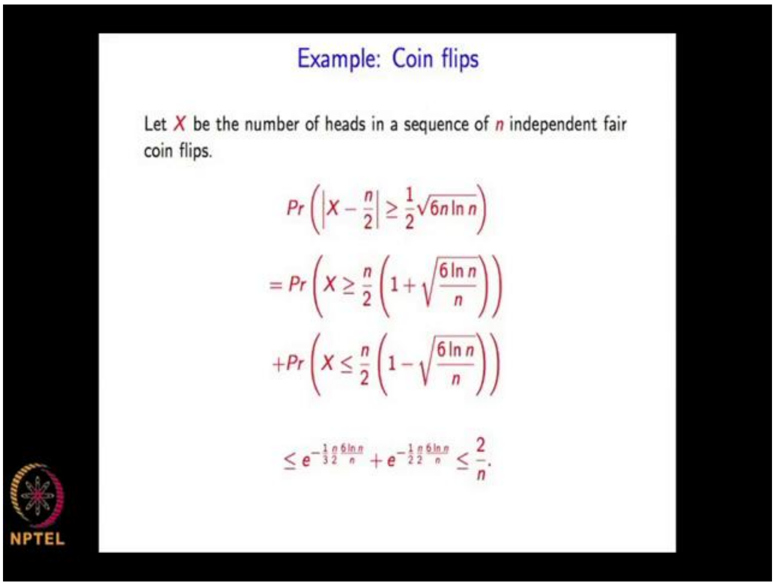
We not do it in detail, but a similar bound work on the other direction. So, Now, you ask what is the probability that the random variable is delta y below the expectation. So, what is probability that instead of getting the expectation I go to value that is delta y below that, and you get basically the same bound only here you have 2 instead of 3 technicalities. You do it the same way we already computed the expectation. So, we use the same bound only thing is now we have to use a t in order to minimize we have to use a t that is smaller than 0 because we switch the order of the inequality, but this is exactly two for this delta that is between 0 and 1 which is what you can plug here. So, let us give you the box.

Now both of this one, so let us take the simplest one. So, assume that the I flip a coin a flip a coin half half coin n time, and ask what is the probability that the number of heads. So, what is probability that the number of heads deviate from the expectation by more than this value half square root 6 n log n. So, now, we plug it into the Chernoff bound. So, to plug it we want to here we have to write it a little bit different we want to as suppose with x is greater than the expectation plus the deviation, and we look for the one going up a above that, or below that is and we get that the probability is two over n.

Now, how go to this bound. Now, you know. So, many intuitions behind now do not just take the symbols. So, how good this well. So, if I flip the coin, n time and we are saying now it is (Refer Time: 37:02) random variable does not to be flip the coin, but flip a coin is a good example how tight I could actually expect the actual result to be half a (Refer Time: 37:12) should expect it to be from the expectation.

Always, definitely very unlikely to be exactly the expectation; the probability of getting exactly half is very small. So, well you actually expect it to be, where most of the mass well that is a standard deviation is. So, I cannot expect to prove a bound, but is better than plus minus 1 or 2 standard deviation, because it is just now, whose high probability you know it is somehow distributed in plus minus 1 or 2 standard deviation you can (Refer Time: 38:00) the bound in that.

(Refer Slide Time: 38:09)



Example: Coin flips

Let X be the number of heads in a sequence of n independent fair coin flips.

$$Pr\left(\left|X-\frac{n}{2}\right|\geq\frac{1}{2}\sqrt{6n\ln n}\right)$$

$$=Pr\left(X\geq\frac{n}{2}\left(1+\sqrt{\frac{6\ln n}{n}}\right)\right)$$

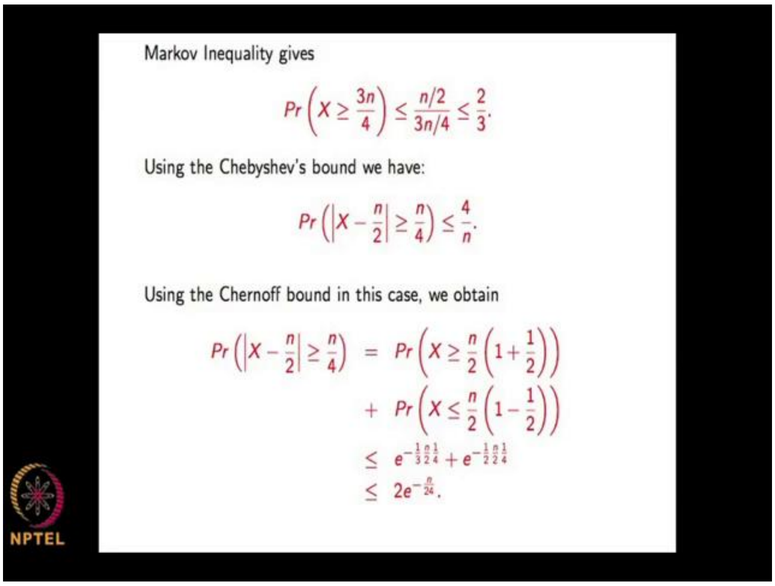
$$+Pr\left(X\leq\frac{n}{2}\left(1-\sqrt{\frac{6\ln n}{n}}\right)\right)$$

$$\leq e^{-\frac{1}{3}\frac{n}{2}\frac{6\ln n}{n}}+e^{-\frac{1}{2}\frac{n}{2}\frac{6\ln n}{n}}\leq\frac{2}{n}.$$

What is the standard deviation here? The standard deviation in here is square root of n. Call some constant theorem. Then the lon here is what give us this whole probability. So, I said that because often you see that you know you trying to prove something, and then you I can you I can it does not work and sometime you just steps.

I should, I can look at it and say I cannot work what are I try to prove which is not true. So, random variable you know expect it if we have more and more observation we expect the average to get closer and closer to the expectation. Where we never be the expectation and it is going to be somewhere, you know plus minus standard deviation, 2 standard deviation. You will get the theorem. So, now, let us do this comparison between the 3 bounds.

(Refer Slide Time: 39:16)



Markov Inequality gives

$$Pr\left(X \geq \frac{3n}{4}\right) \leq \frac{n/2}{3n/4} \leq \frac{2}{3}.$$

Using the Chebyshev's bound we have:

$$Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{4}{n}.$$

Using the Chernoff bound in this case, we obtain

$$\begin{aligned} Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) &= Pr\left(X \geq \frac{n}{2}\left(1 + \frac{1}{2}\right)\right) \\ &+ Pr\left(X \leq \frac{n}{2}\left(1 - \frac{1}{2}\right)\right) \\ &\leq e^{-\frac{1}{3}\frac{n}{2}\frac{1}{4}} + e^{-\frac{1}{2}\frac{n}{2}\frac{1}{4}} \\ &\leq 2e^{-\frac{n}{24}}. \end{aligned}$$

You now to get any strong results with the weak bond we have to ask much weaker question.

So, let us ask what is the probability that is; when I flip a coin n time I get 3 quarters heads that is really unlucky. So, if I plug it into the Chernoff Bound to Markov inequality there are probability that is I gets three quarter n helps well probability is bounded by the expectation time the value here the a, and we get that the probability is bounded by two-third is really nothing, is not bound here and particular what is real we seen is that the intuition is that is f if we flip more and more coins, if we do more and more experiment we expect is somehow to converge to get closer and closer to the expectation and here we gets bound that as nothing to do with the member of experiment with it. So, let us Markov Inequality where would if it channel Chebyshevs.

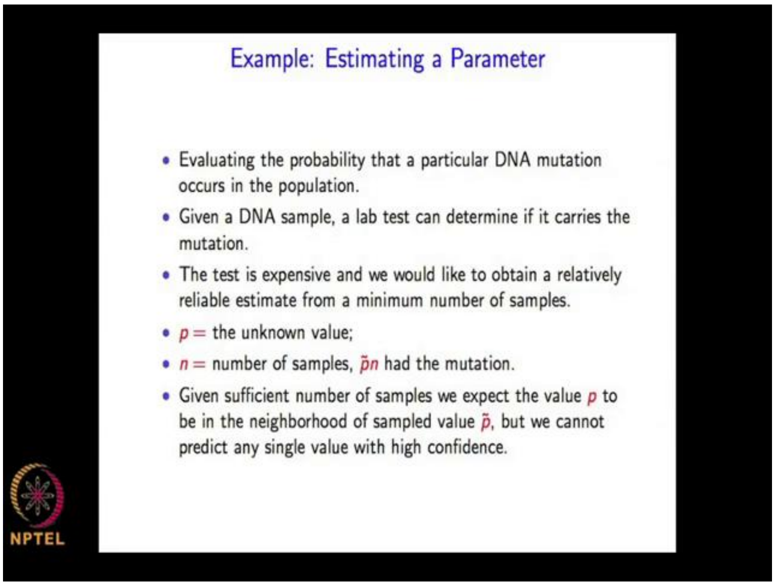
So, the probability that we deviate by from the expectation by n over 4, now, if you have remembered Chebyshevs we have to that is bounded by variance divided by this square. So, the variance here is half time, half time n and we divided by this square. So, we get n square then here and appear we cancel. So, we get at least something that goes down to 0 who is n is n goes to infinite, but it goes down the linearly is n. So, now, we look at Chernoff Bound which we just proven. So, the probability x deviates from expectation by n over 2. So, let us probably that x is greater than it is expectation 1 plus half or it is smaller than expectation by 1 minus half. Now, we plug it here and remember that is bound that we move, is the expectation time the deviation square.

The expectation here is n over 2, the deviation is half. So, it is quarter here, then we have some coefficient here and what we get here is bound that goes down to 0 exponentially in n. So, going for one to the other we started with the bound that it you can go down to 0 as n goes to infinity. Then Chebyshevs gave as a bound that goes down to 0 as n goes to infinity, but it goes down as one over n. That is linear and then we get the bound that goes down to 0 whose n exponential in n.

So, that is significant jump between these bounds, with the beauties is that all build on Markov inequality that is. So, weak is. So, weak on it is own. So, that is the basic bound Chebyshevs. The Chebyshevs bound and comparison of the C hernoff bound. So, that is in particular when use it in algorithm what was it say he says that if I want to get something that is you know the exponential is small or If I want to show that the algorithm you know work fast same as probability one over n, then I just need log n experiments or log n values instead of n values, that is a huge jump when it was (Refer Time: 43:38) you know execution of algorithms, we will see example later.

So, the simple application of all these is estimating the parameter, and the caution is actually note question (Refer Time: 41:01) a note question computer science questions in statics, but it is give as the right idea of what happening here.

(Refer Slide Time: 44:03)



Example: Estimating a Parameter

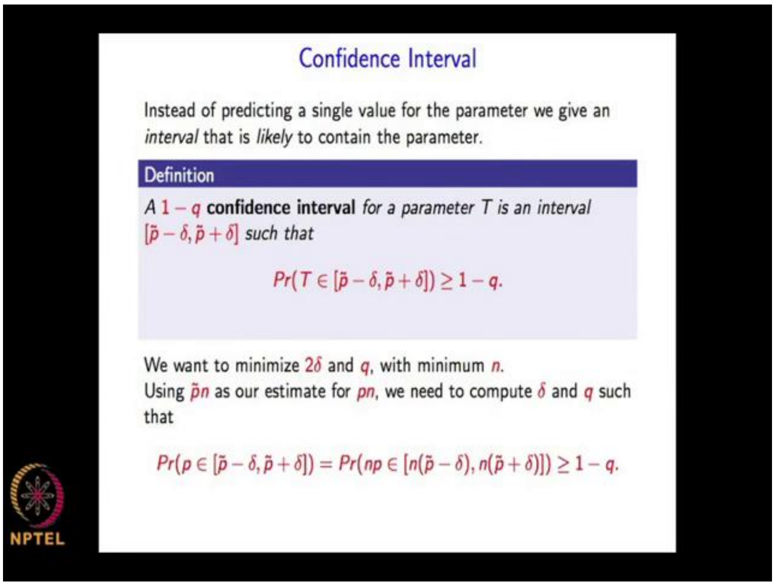
- Evaluating the probability that a particular DNA mutation occurs in the population.
- Given a DNA sample, a lab test can determine if it carries the mutation.
- The test is expensive and we would like to obtain a relatively reliable estimate from a minimum number of samples.
- $p =$ the unknown value;
- $n =$ number of samples, $\tilde{p}n$ had the mutation.
- Given sufficient number of samples we expect the value p to be in the neighborhood of sampled value $\tilde{p}$, but we cannot predict any single value with high confidence.

So, assume that there is some p value and on. So, the story here is I want to know the fraction of the population that has particular mutation. If I am getting sample for someone in it black sample some where we can test and I can figure out if he or she has to be mutation or not. Well I could go in a test all the population of India, this may take long time. So, the question is you know how many samples I need in order to get the very good estimate for the whole population.

And the same question of course, is when you try to predict the lecture is as an you know all the other things and of course, is most people other than scientist like you, do not understand is in order to estimates the value in a population. Where you should do not care about the size of the population. The sample size as we see as nothing to do with the size of the population, which is always very (Refer Time: 45:24).

So, let predicting the election is at in India and in Singapore takes the same amount of, the same 1400 sample assuming that you give me actually from sample. Whatever I know it is much more complicated than that. So, so you have to evaluate you want to estimate the value that is unknown. So, we have in what is called classic statistic. So, p is some unknown only God knows that and it tells us. So, we are only estimating it we will never actually know the real value. We will take n samples in p till the n will be the fraction of the samples that, had what we looking for, the mutation. So, give a sufficient number of samples we expect the value p, to be in the neighborhood of our estimated value p tilde up.

(Refer Slide Time: 46:40)



Confidence Interval

Instead of predicting a single value for the parameter we give an *interval* that is *likely* to contain the parameter.

Definition

A $1-q$ **confidence interval** *for a parameter T is an interval* $[\tilde{p}-\delta, \tilde{p}+\delta]$ *such that*

$$Pr(T \in [\tilde{p}-\delta, \tilde{p}+\delta]) \geq 1-q.$$

We want to minimize 2δ and q, with minimum n.
Using $\tilde{p}n$ as our estimate for pn, we need to compute δ and q such that

$$Pr(p \in [\tilde{p}-\delta, \tilde{p}+\delta]) = Pr(np \in [n(\tilde{p}-\delta), n(\tilde{p}+\delta)]) \geq 1-q.$$

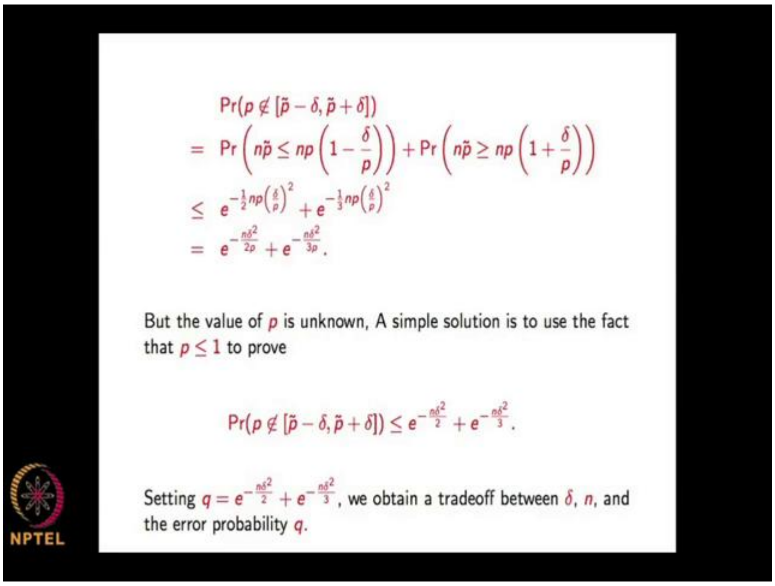
Well one way we can use the Chernoff Bound is to estimate, how close is p tilde to p. Is the function of the number of samples? In computer science usually refer to this as we often do it now machine learning, (Refer Time: 46:45) of that we answer this kind of questions, and the two ways to estimate the value was in statics and machine learning. One of them is go what is go to point estimate the other one is a interval estimate. Now point estimates is very (Refer Time: 47:07) because most the point estimates it just coincidence that is give sample give this particular value is a maximum, but you know the value can be in the neighborhood of this. In particular if you think about.

So, in much more will have a (Refer Time: 47:29) to get into this we can do in some other, over much more reliable and meaning full estimate is what is called a Confidence Interval. Confidence Interval is I will give you. So, confidence interval for a particular value t is I give a range; such that is very high probability the value is in inside that range that is something very confusing here.

T is this value that is, return you know God knows about it. But it is a fixed value a random variable is in fact the interval, goes the interval is what we get from a sample, the values that we get from the sample. We say well if we got p tilde, is a expectation of the sample we go delta plus and the delta minus and we say well we expect the real value to be inside this interval. So, this is actually random variable this is a fix value in nature. So, now I want to minimize.

So, the real thing here is you have these 3 parameters you want to play with; you want to minimize the intervals. So, that you have a meaningful estimates. You want to minimize the arrow probability that is q, and you want to minimize the number of sample because that is what you pay for. That is what the work and the question is; you know how do you get the tradeoff between them? So, you want the probability that is the interval include the real value of p, which is the same as they number of you know positive observation you had minus delta, plus delta you want this interval to include n time p in order to be use high probability.

(Refer Slide Time: 49:44)



$$\begin{aligned} & \Pr(p \notin [\tilde{p} - \delta, \tilde{p} + \delta]) \\ = \; & \Pr\left(n\tilde{p} \leq np\left(1 - \frac{\delta}{p}\right)\right) + \Pr\left(n\tilde{p} \geq np\left(1 + \frac{\delta}{p}\right)\right) \\ \leq \; & e^{-\frac{1}{2}np\left(\frac{\delta}{p}\right)^2} + e^{-\frac{1}{3}np\left(\frac{\delta}{p}\right)^2} \\ = \; & e^{-\frac{n\delta^2}{2p}} + e^{-\frac{n\delta^2}{3p}}. \end{aligned}$$

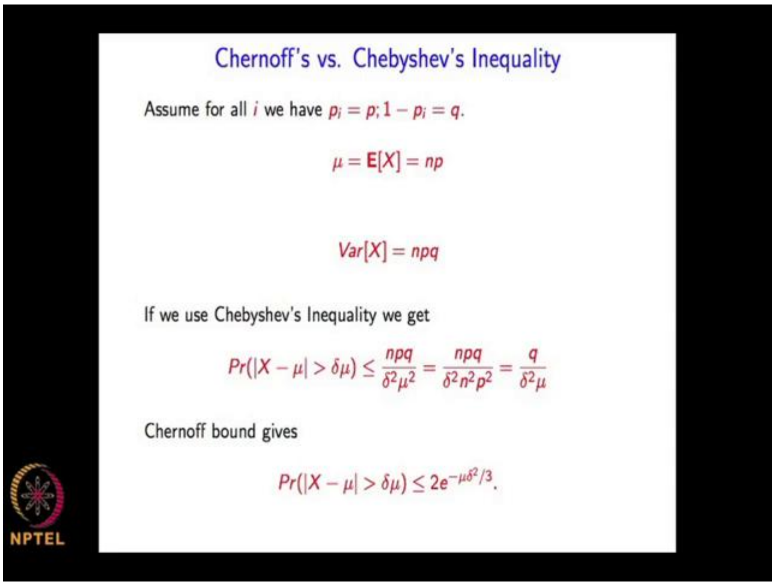
But the value of p is unknown, A simple solution is to use the fact that $p \leq 1$ to prove

$$\Pr(p \notin [\tilde{p} - \delta, \tilde{p} + \delta]) \leq e^{-\frac{n\delta^2}{2}} + e^{-\frac{n\delta^2}{3}}.$$

Setting $q = e^{-\frac{n\delta^2}{2}} + e^{-\frac{n\delta^2}{3}}$, we obtain a tradeoff between δ, n, and the error probability q.

So, we have to look it is above that and below that, is and I want to get the detail of the calculation, but basically you gets, you plug it into an Chernoff Bound and you gets a basically the relation between you want these to be bounded by q, you get n and delta here. So, you get the full relation between this (Refer Time: 50:02). So, that is one simple application of a Chernoff Bound that gives you very good values and actually it give you very (Refer Time: 50:14) estimate and again we are playing with delta. Who is q, Who is n the size of population does not appear in this whole computation.

(Refer Slide Time: 50:36)



Chernoff's vs. Chebyshev's Inequality

Assume for all i we have $p_i = p; 1 - p_i = q$.

$$\mu = \mathbf{E}[X] = np$$

$$Var[X] = npq$$

If we use Chebyshev's Inequality we get

$$Pr(|X - \mu| > \delta\mu) \leq \frac{npq}{\delta^2\mu^2} = \frac{npq}{\delta^2 n^2 p^2} = \frac{q}{\delta^2\mu}$$

Chernoff bound gives

$$Pr(|X - \mu| > \delta\mu) \leq 2e^{-\mu\delta^2/3}.$$

Let me. So, finish this. So, let me just (Refer Time: 50:40) about what we said before because again this is important. So, assume that we have let us, I just want to give you the more precise comparison on between the Chernoff and the Chebyshevs inequality. So, assume that we have, this assume to I think easier that which is sum n a random variables which probability p for success, assume random variable probability p for success. So, the expectation of the sum of n variable a this n 0, 1 random variable is np the variance is npq and if you know all these.

So, if we use chebyshev you gets the probability that you deviate from the expectation by delta time the expectation, is bounded by the variance divided by this square of the deviation, and if you plug the I think you get that is basically q, divided by delta square mu. And that is important one is delta square mu and now actually do the Chernoff bound. So, the probability that you deviate from by mu, from mu by delta is bounded that e to the minus mu. Delta square over 3.

So, that this delta square mu, that pr involves cases is to say one only here at the pr is 1 over that then here it is minus that in the exponent. So, the real thing that is go down to the real measure, of convergence is in the expectation time the variation square time they deviation square. Where is that? (Refer Time: 52:39).

Student: (Refer Time: 52:41).

I do not hear.

Student: Which Chernoff Bound as the best bound you can have say for (Refer Time: 52:49).

No. So; as usual there is a tradeoff between general a tools and the best you can get. So, if you give me a particular problem, you know I can probably squeeze some of better bound, because to get the bound we did also of simplification here, but the.

Student: for the example of (Refer Time: 53:15) how can you implement?

For example of the point estimate.

Student: Point need of the co-ordination from the expectation.

Yeah.

Student: So, do for Chernoff bounds.

Yeah.

Student: To get a better upper bound using the specific information for the expectation.

Well. So, it depends on what information you have, but in you would not get much better. So, they will see next time, that if (Refer Time: 53:42) the Chernoff Bound for very simple case of you know two value and the variable converge it gets it looks very, very similar to answer the normal distribution. So, beyond the fact that it is you know that is the question of convergence now a distribution we do it with respect to particular n. We are getting very close to the limits.

So, but again when we develop the bounds, we gave up a few we gave a little bit in getting the you know, easier to work bounds, but the difference is almost significance. So, the exponent would not change, that is may be some constant even there, but you know asymptotic that is the best you can get at least for the case of 0, 1 random variable.

Student: (Refer Time: 54:42).

Yeah.

Student: This is not a best (Refer Time: 54:43) in terms of (Refer Time: 54:44).

I do not hear you.

Student: can you prove that this is one of the best you can get in terms of asymptotic.

Again, so you can show it for the case of a 0, 1 random variable you can show that the asymptotics you get very close to the normal which is the limit.

Student: (Refer Time: 55:02).

Ah not directly but yeah that is part of the reason, but they again, if all of you if you if you if you have n identically distributed random variable 0, 1, then yeah then it could converge the very basic you show that you converge to a normal.

Student: Using.

Yeah. So, you show that it is. So, basically you get you know in the limit you are in the optimal, but if you start with random variable that of different, then I do not think you could get, you can show the directly that is optimal. So, there might be some twists that you can do more use the particular distribution, but then with the all idea of a building this kind of tools is that you do not have to work, you know the details for which distribution, but you get something that works in general or yeah.

Student: based on this question if we know if we have some trias, then how we can use them to say.

Something about bounding; if we have a bias.

Student: Trias.

Student: (Refer Time: 56:19).

Or if you now you want to do bias here that are, yeah you can do that is, but that require some stronger tools. So, one way, keep your question for the next hour, I mean for tomorrow. So, one way to look at bias is in terms of (Refer Time: 56:45) here that is one way to look at it.