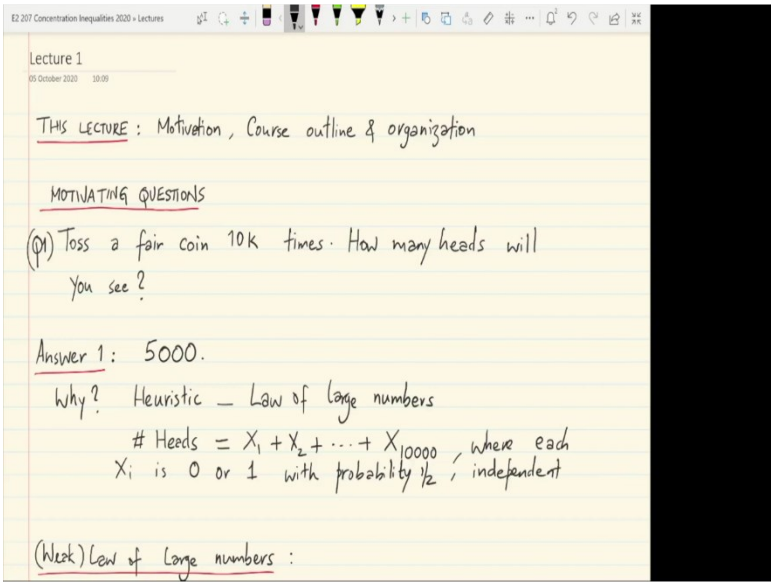
Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture – 01
Why study concentration inequalities?

(Refer Slide Time: 00:20)



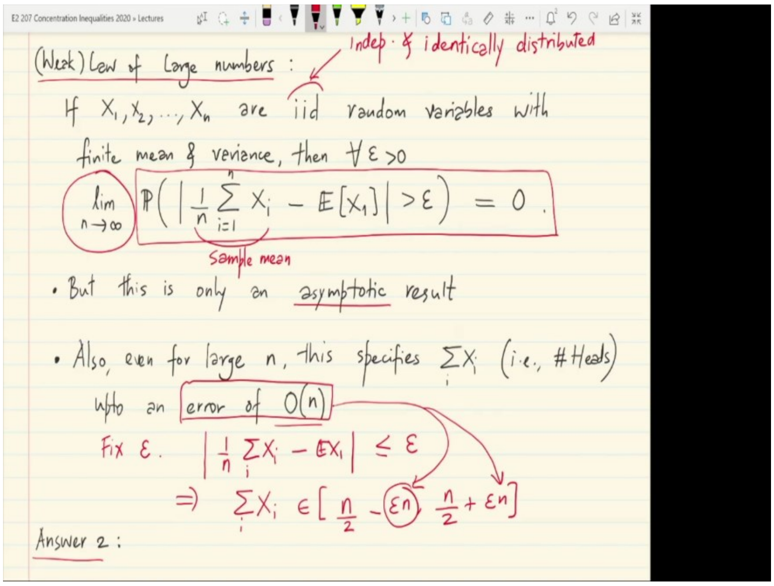
Hi all. In this lecture, we will study some motivating examples for the course on Concentration Inequality. And we will end with some observations as to what the course will contain, and provide a high level overview of the organization and content of the course.

So, let us start with a very basic motivating question, ok. So, suppose you toss a fair coin independently 10000 times, how many heads will you see? Ok. So, at a first cut you may be tempted to answer that you expect to see 5000 heads, that is correct you indeed expect to see on an average 5000 heads, but then there are several levels at which this question can be answered.

So, let us first begin/asking the question what is the reasoning behind 5000, ok. So, why is this? This is based on the heuristic that as you toss more and more times the number of heads observed should essentially become closer and closer to the expected value or the average number of heads from tossing a single coin, ok. So, this is called the law of large numbers, ok.

So, one can actually represent mathematically the number of heads that you will actually get as the sum X 1 + X 2 all the way up to X 10000, ok, where each random variable X i is 0 or 1 with probability half, and these are independent, ok. So, what is this law of large numbers?

(Refer Slide Time: 02:19)



So, standard probability theory tells you that there is let us say this weak law of large numbers which basically states that if X 1, X 2, and so on up to X n are independent and identically distributed random variables, so iid will always be used to mean independent and identically distributed.

So, if the X i are random variables with let us say finite mean and variance then for any tolerance ε we have that the probability that the sample mean of these X i's converges to the expected value of any one of them. Recall that they are all identically distributed anyway.

So, the probability that the sample mean deviates from the true expectation/a quantity more than ε tends to 0 as more and more random variables are added to the mix, ok, so as n increases. But however, this is only an asymptotic result, ok. So, note that there is limit here. So, it is not a priori easy to say at what end and how small this probability becomes, ok. So, at what end this probability starts becoming very small is not indicated/this result and that is the reason it is called an asymptotic result, ok.

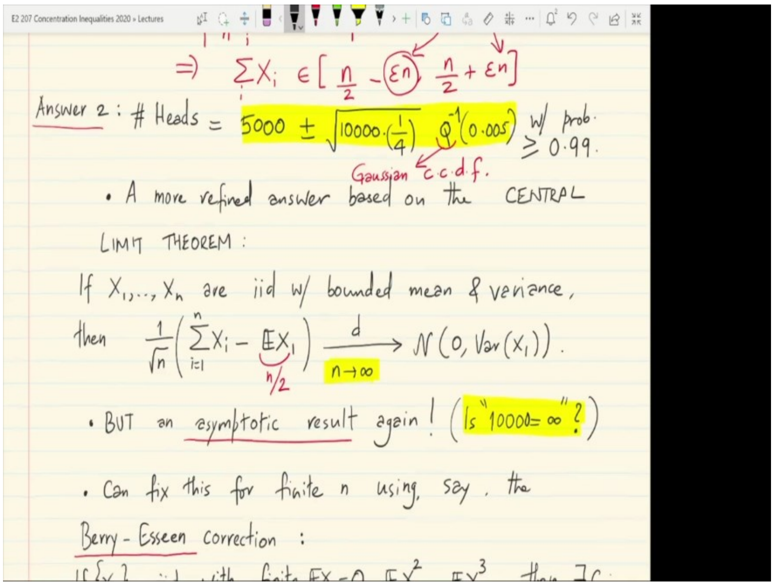
So, even so, there is another kind of looseness with this law of large numbers, I am trying to answer this question, which is the fact that even if n is large, let us say n is large enough such that the probability above has become exactly 0, ok. So, let us say that you have made n so large as to drive this probability to approximately 0, ok.

What we have from this statement is that you can fix any ε, ok. So, you can fix, so fix an ε and this will tell you that you know 1/n summation i X i – E(X 1), is < = ε, ok. So, at a very first order level of approximation you will have that this quantity is < = ε.

So, in other words this means that the total number of heads which is the sum i X i is basically within n times expected value of X 1 which is half, ok - ε n and the upper end of this interval is n/2 + ε n, ok. So, this means that the amount of approximation in reporting or estimating the sum X i or the number of heads is basically an order n amount, ok.

So, the order n amount is basically this ε n here, ok. So, this specifies the number of heads up to an error of order n, even if you assume that it is not asymptotic, ok. In some strange way, suppose it were non-asymptotic, you would still get an order n error here, ok.

(Refer Slide Time: 05:30)



The second type of answers that you can give to this question is by using something more refined, ok. So, you can probably for instance say the following. You can say that the number of heads is going to be let us say 5000 which is the mean value, which is what we estimated in the first answer + or - a certain range.

In this case, this is the √the number of times you are tossing this coin that is 10000, into the variance of each toss, which happens to be 1/4. And there is this function Q inverse of 0.005. So, the answer here is that the number of heads is 5000 + or - this number here, this deviation with a probability let us say at least 0.99, ok. So, with 99 percent probability the number of heads is bound to be within this range, ok.

So, what is Q here? If you have not seen this before, Q is just the Gaussian tail probability function, the Gaussian cumulative complementary cumulative distribution function, ok. So, Q is the Gaussian CCDF and Q inverse is its inverse, inverse function. So, Q of any number is the probability that a standard normal Gaussian random variable is > or = that number, and Q inverse of any number is the inverse of the Q function, ok.

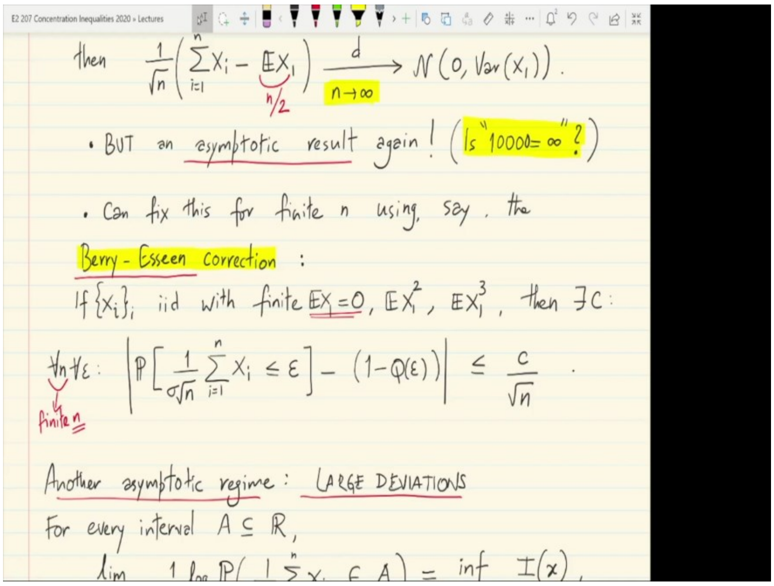
So, this is clearly a more refined answer. So, this 5000 + - something, is clearly a more refined answer, which is actually based on what is called the central limit theorem which is another asymptotic result. So, what is the central limit theorem here? So, one version of the central limit theorem which is a very famous result in probability is as follows.

So, if X 1 through X n are iid with bounded mean and variance let us say, then if you take the quantity summation of X i, which is the number of heads in our case, and center it/its mean which is n/2 here, ok. And divide this deviation/a factor √n, then as n increases the this quantity this random variable converges in distribution to a standard normal random to a normal random variable with mean 0 and variance = the variance of X 1, each of these random variables, ok.

So, using this one can try to also estimate or give some estimates of deviations. But again this is an asymptotic result because you see there is a limit here, ok because of the limit here this is again an asymptotic result, prompting questions like you know is 1000 = infinity, ok.

So, in some sense is sorry this is 10000, so is 10000 = infinity, is the number of heads that we have tossed, is the number of coins that we have tossed sufficient to guarantee that these statement of the central limit theorem kicks in or not, ok. So, this is not explicitly indicated/the theorem, ok. So, we have come this far, and this gives somewhat sharper estimates than the answer number 1. One can hope to fix the fact that this is an asymptotic result/using what is called Berry-Esseen correction, ok.

(Refer Slide Time: 09:08)



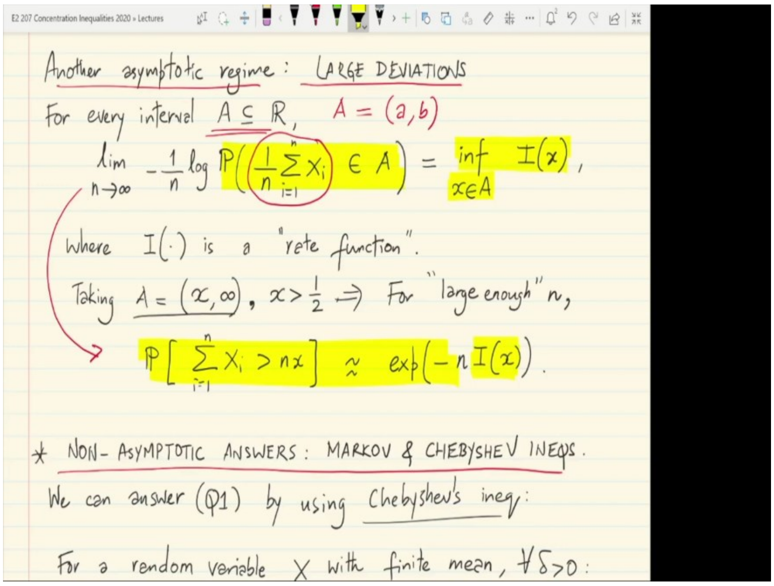
So, Berry-Esseen correction is essentially a fix to make the standard central limit theorem non-asymptotic, ok. So, one version of a Berry-Esseen type correction is as follows. If all these random variables are iid with finite expectation, let us assume that they are 0 mean, finite variance and finite third moment then there is some constant such that the following happens.

For any number of tosses note that this is non-asymptotic, ok. So, at any finite n and any value of deviation ε, the probability that the scaled number of heads, essentially scaled/1/√n is < a certain number. So, in reality what would you expect this to be? If the central limit theorem were perfect you would expect this to be the cumulative distribution function of a standard normal which is 1 - Q of ε.

And so, the Berry-Esseen correction basically tells you that these two quantities differ/a quantity, which is no more than constant/√ n, ok. So, the larger now is if the number of heads you toss the tail probability of the scaled sum versus the benchmark, which is the corresponding tail probability for a standard Gaussian differs only/a vanishing quantity that goes particularly as constant/√ n, ok.

So, you can use this to correct your estimate for a number of heads a little better. And this has been this is widely used in probability theory. But however, we will see that more generic and simpler to understand techniques also give us similar answers, ok. So, that is sort of going to be the main punch line of this course. You do not need to study specifically what happens with Gaussian random variables or you do not need to try to match things with Gaussian random variables, but you can develop generic techniques.

(Refer Slide Time: 11:17)



So, another comment in passing is that there is another asymptotic regime different than the central limit theorem regime which is called large deviation. So, recall that the characteristic feature of the central limit theorem is that the sum of n terms is being divided/something like √n, ok.

In the large deviations regime, it is more like dividing the sum/the total number of summands itself. So, it is the sample mean, regime. It studies the fluctuation of the sample mean. So, here is a typical large deviation statements specialized to our case of independent coins being tossed. So, pick your favourite interval A, ok. Let A be an interval in R, ok. So, let us say A is of the form a , b, ok.

So, it is the set of all real numbers between a and b. The intervals can either be open or closed, it does not matter really for this example. What it says is that if you take the probability that the sample mean of n coin tosses, the average proportion, that the proportion of heads in those tosses belongs to this interval A, if you take the probability that this proportion belongs to A, take the log of that probability and divide/1/n with a - sign, ok.

So, that is going to be a non-negative number finally. As n become larger and larger, this quantity tends to a very specific limit which is the, which can be expressed as the solution to another variational problem. The minimum of a certain function I on that interval A, ok.

So, think of this function I called a rete function, defined on the entire real line and if you are interested in the probability of this sample mean falling in any interval for a large enough value of n, all you have to do is go to that interval and solve the minimization of I over that interval and you will basically get this exponent, ok.

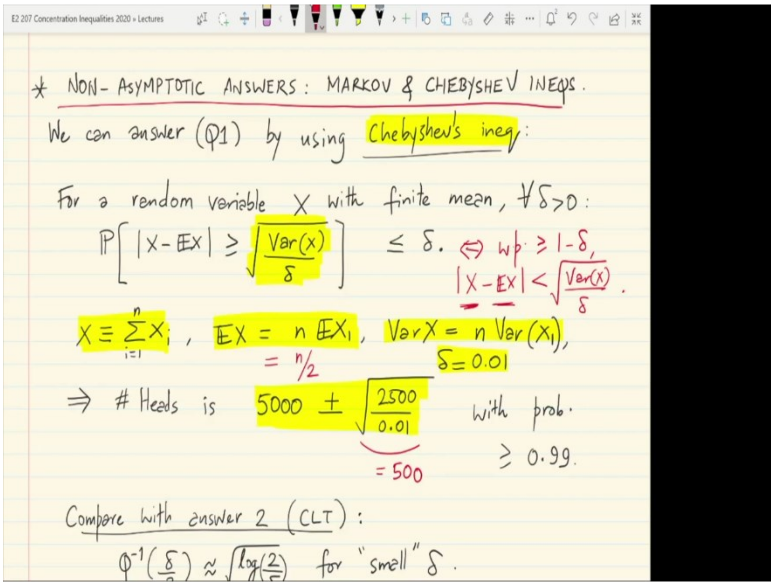
So, what this means in other words is that if you take A as the interval x to infinity let us say where x is larger than half, ok. So, we need x to be larger than the expected value which is half in this case of each random variable, then what happens is something nice. For large enough n the probability that summation X i larger than a number n times x approximately falls off exponentially in n with the constant in the exponent being given/this the rate function here, ok, at x.

So, this is just/manipulating the general large deviations inequality above. So, the general large deviations limiting result tells you that - 1/n log of this probability is roughly going to be the infimum of X in A of I of x, ok and it turns out that you can show that I of x must be a convex function and it always is non-decreasing to the right of the expected value.

So, the minimum of I of x the minimum of the function I in the interval x to infinity, where x is to the right of the actual mean, mean of x 1, which is half that answer must be I of x itself. So, you basically get a slightly stronger in some sense statement of the weak law of large numbers, in the sense that you get this decay rate, exact decay rate of the probability of deviating from the sample mean, and this is an exponentially decaying rate.

However, this is still asymptotic. So, there is no indication provided of at what n this limit really becomes close, ok. So, leaving this aside let us proceed to something, which is actually going to be the main motivation for undertaking the study of concentration inequalities.

(Refer Slide Time: 15:26)



It is the fact that one can actually get rather concrete non-asymptotic answers using things simple inequalities or statements like Markov's and Chebyshev's inequality, ok. So, for instance, we can answer question 1, ok which is with the question of how many heads do you expect to see in 10000 tosses/just using what is called Chebyshev's inequality, ok.

So, Chebyshev's inequality says that for a random variable X let us say with finite mean, for any value of $\delta > 0$, typically δ is a probability number between 0 and 1, we have that the probability of X - its expectation deviating from this number which is √variance X over δ is no more than δ, ok.

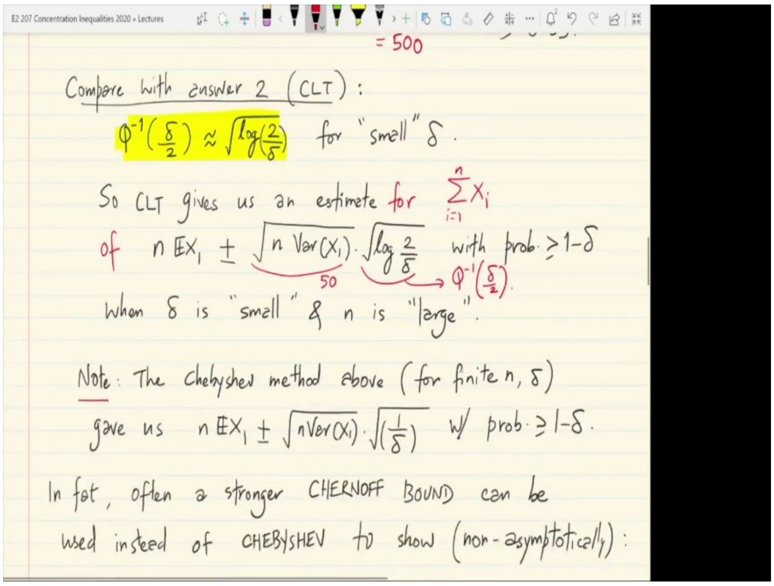
So, the smaller the δ the larger the bound on X - EX, ok. So, this just means if you turn it on it head that with probability at least 1 - δ, the quantity X - expected value of X is always < √ variance X divided/δ, ok. So, this gives you a high probability guarantee on X deviating from its own mean, ok this is one of the most, this is arguably one of the simplest concentration inequalities.

And if you now specialize this to X being the sum of the first n coin tosses, where each X i is either 0 or 1, probability half independently. The expected value of X simply becomes n times the expected value of every coin toss which is n/2, ok. So, this is n/2, ok.

Likewise the variance of X, which is the sum of n independent random variables you can also show that this is equal to n times, this is the sum of variances really. So, the sum the variance of a sum of independent random variables is just the sum of the individual variances.

And let us say you set δ = 0.01 and apply Chebyshev's inequality to this setting you will basically get that with probability at least 0.99 which is 1 - δ, the number of heads is this quantity which is 5000 + - 2500/0.01, yeah inside the √, ok and this evaluates to 500, ok. So, you get at least a concrete answer which is guaranteed to be correct with probability at least 99 percent if you toss exactly 10000 coins, ok.

(Refer Slide Time: 18:20)



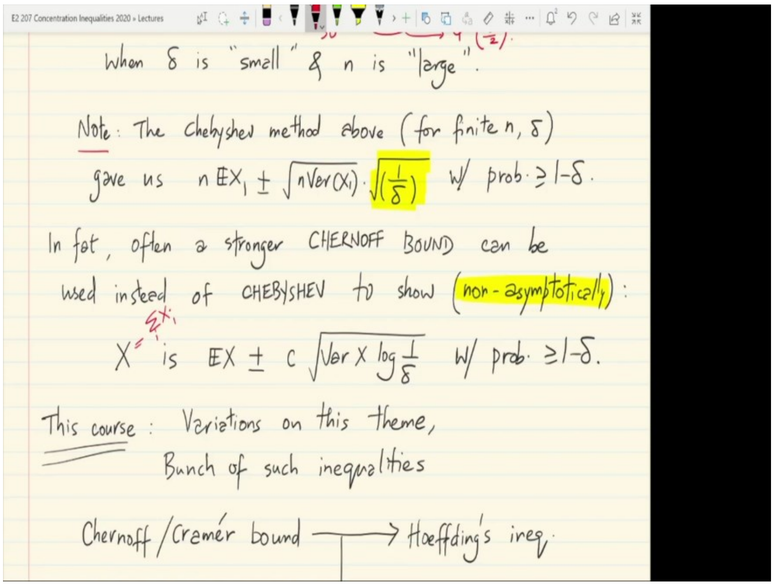
So, compare this with, let us compare, let us take a moment to compare this with the CLT inspired answer, the central limit theorem guided answer, which gave us something similar, but/making sort of the hand wavy approximation that 10000 is close to infinity, ok. So, what did answer 2 gave you? Maybe it is time to go back and look at answer 2.

So, answer 2 basically gave you 5000 + - you know √2500, which is 50 times a Q inverse of 0.005, ok. So, what is Q inverse 0.005? If you actually look a little carefully at the Q inverse function, it is not hard to see that show that Q inverse of a very small number is approximately √log of inverse of that number, ok.

So, for δ very very small Q inverse of δ/2 is roughly √log of 1/δ/2, which is √log of 2 over δ. So, the CLT gives us an estimate for you know for the summation i = 1 to n X i which is a total number of heads of n times expected value X 1, which is the expected value of the sum + - √n times variance of X 1 that is exactly the central limit theorem correction, ok.

You have to divide/σ into √n, where σ is the standard deviation of each random variable, ok into a √log 2/δ, ok. So, that √2 log/δ is the approximation being used here to Q inverse δ/2, ok, with probability at least 1 - δ. So, note that this number here, the √n into variance, which is 50 in this case is being multiplied with √log 2/δ.

(Refer Slide Time: 20:30)



Whereas, in contrast the finite sample Chebyshev inequality above, gave us a valid inequality for any finite n and for any δ, any probability δ of order n times expected value X 1 + √ n times variance X 1 into √ 1/δ, ok. The √ 1/δ is simply the √ 1/δ here, ok.

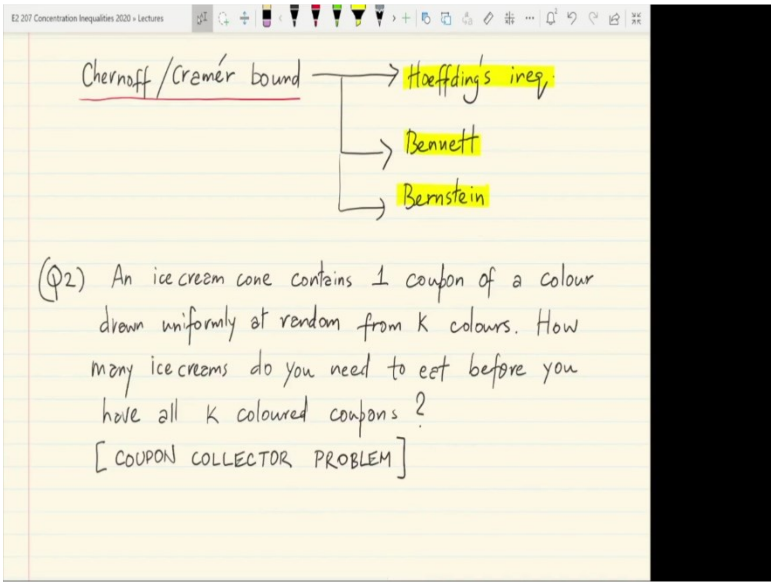
So, that is the √ 1/δ, with probability at least 1 - δ. So, there is this sort of loosening here in that the order of deviation here that Chebyshev gives you a √ 1/δ as opposed to √ log of 1/δ, ok. So, this is a arguably a big factor improvement in the CLT, which is at a different level weaker because of being asymptotic.

But, there is nothing to despair here. In fact, we will see that there is often a stronger method called the Chernoff bound compared to the Chebyshev bound that can be used instead of Chebyshev to show in a non-asymptotic fashion for finite samples n that you can get the same order of magnitude as what CLT gives you, ok.

So, you will basically get with X being the sum of n coin tosses, you will basically get the actual expected mean + - sum constant times the √n times variance X 1 into log of 1/δ inside the √, ok. So, you get a big improvement of going from 1/δ to log 1/δ, ok.

So, that is where sort of Chernoff really matter. It gives you in some sense an exponential rate of decay of the probability of the sample mean deviating, which if inverted gives you a logarithmic dependence on the reciprocal of the target failure probability.

(Refer Slide Time: 22:25)



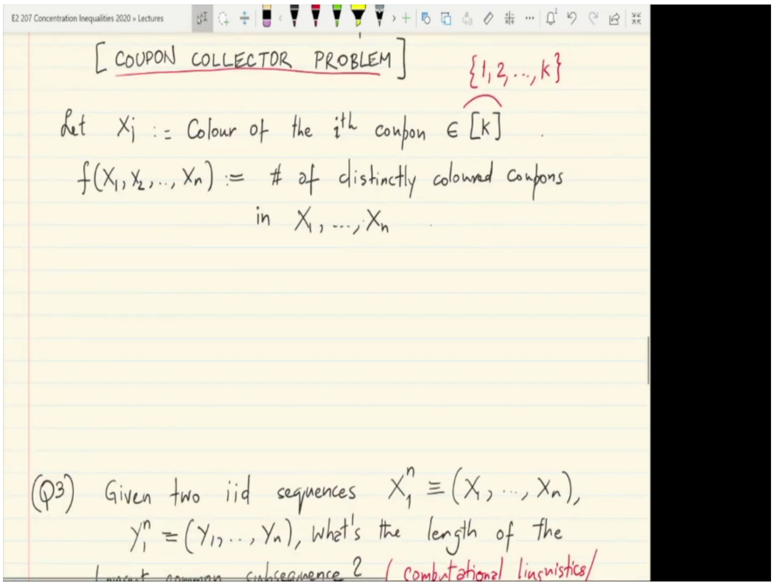
So, this course is basically going to be presenting a lot of variations on this theme for both independent as well as non-independent random variables. And they are going to be a bunch of such derived inequalities which build upon basic Chebyshev or Chernoff methods, ok.

So, at the heart of all these methods is going to be what is called the Chernoff Cramer bound or the Chernoff Cramer technique, which will which have used or instantiated appropriately gives very useful inequality such that Hoeff, such as Hoeffding's inequality widely used in areas like a learning theory and statistical learning theory.

It gives inequalities like Bennett's inequality and Bernstein's inequality which is often used as a tighter version of Hoeffding's inequality, ok. So, with this let us move to the second motivating example. The second motivating example is the following question, which is often called the coupon collector problem.

So, let us say every day you go and buy an ice cream cone and hidden within it there is one coupon, every ice cream cone contains one coupon of a uniformly random colour, ok. So, let us say there are K total colours of which coupons are made in each time every ice cream that you buy contains one randomly coloured coupon, ok hidden inside in it and you start collecting these coupons. Now, how many ice creams do you need to eat before you can collect all K coloured coupons? Ok.

(Refer Slide Time: 23:53)

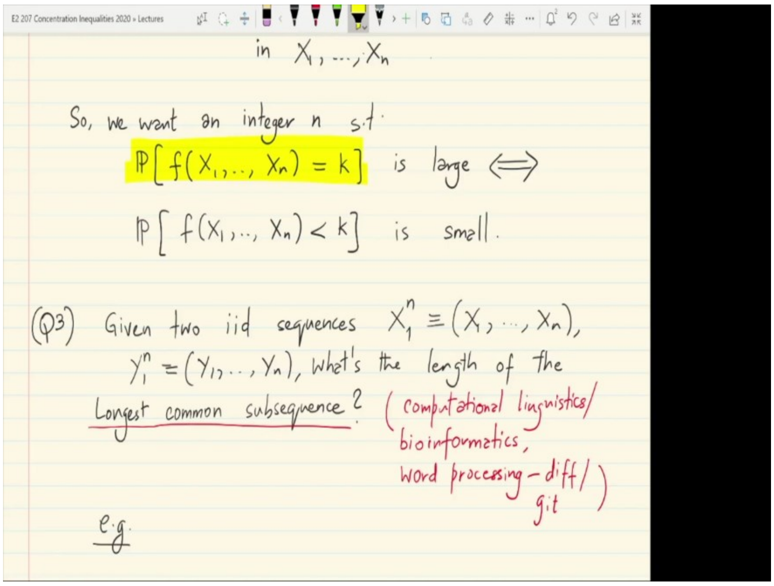


So, this is the well-known Coupon Collector Problem. And just to formalize this again like we did with the coin tossing problem. So, let us denote X i as the colour of the ith coupon, ok that you collect, ok. So, this is a clearly let us index all the colours/1, 2, 3 up to K, let us assume there are K colours.

And so, this is clearly a number between 1 to K, ok. So, this bracket k notation basically denotes the set of integers 1, 2, 3 up to K, ok. So, in addition let us introduce this notation f of X 1, X 2, up to X n. So, suppose you have you have eaten n ice creams and you have n coupons in your hand, f will look at all these X 1, X 2, up to X n the colours of all these coupons and it will return the number of distinctly coloured coupons among the set, ok.

So, let f be the number of distinctly coloured coupons among the coupon colours X 1 through X n, ok.

(Refer Slide Time: 25:24)



So, what is the coupon collector problem? It basically asks you to find with high probability. So, we want an integer n such that if you have collected n coupons then the probability that the number of distinct coloured coupons that you have collected equals the full set of colours K, with high probability.

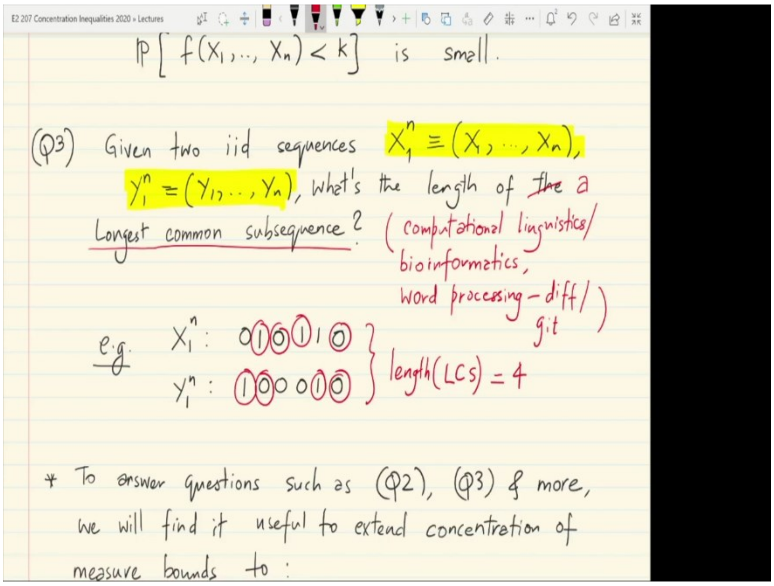
So, this probability should be large or in other words the probability that f of X 1 through X n strictly < K you have not made all colours is quite small, ok. So, at what value of n relative to K do such probabilities start appearing very large or very small respectively? Ok.

So, that is the coupon collector problem. I am not going to get into the basics or the details of this problem. It turns out that you can easily get the expected value of the number of the first time n at which you collect all K coupons, ok. That happens to be in fact, K log K. But let us get into the details of this later in the course if possible, ok.

So, this is just a motivating example to illustrate that you would like to control probabilities of this form, where X 1 up to X n are in this case independent random variables, they are independent colours of coupons that you buy each day, but the f function is something that looks very different than let us say the sum or the sample mean of X 1 thorough X n, ok.

So, it counts something very discrete, it counts the number of distinct colours. So, you want to control the fluctuation of functions like f rather than functions that look like simple sums or averages, ok. So, this is the point to bear in mind from this motivating question.

(Refer Slide Time: 27:28)



The last motivating question is something that occurs in areas of computer science, computational linguistics, sometimes in bioinformatics and word processing and so on. It is the problem of that longest common subsequence. So, this just says that if you picked two, if you made two iids sequences X 1 through X n and Y 1 through Y n, ok.

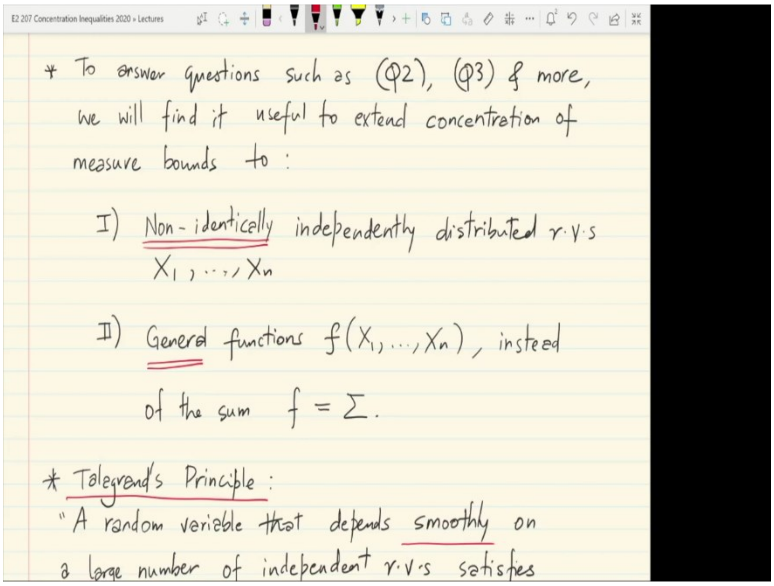
So, each of the X i is independent each of the Y i's is independent from the other random variables here. What is the length of the longest common subsequence between these two? Ok. So, what is the longest common subsequence defined as? I will just describe that using an example. Suppose X 1, so suppose X 1 n is the sequence of let us say 6 0 or 1 letters, so 0 1 0 1 1 0 and let us say X 2 sorry, Y 1 n is another binary string 1 0 0 0 1 0, ok.

The longest common subsequence, the length of a longest common subsequence sorry, I should have said a longest common subsequence, it need not be unique always. So, for instance one common subsequence which occurs in both is 1 0 1 0, similarly you can see the same pattern here 1 0 1 0, ok.

So, there can be any arbitrary number of a symbols in between this common subsequence in when it occurs in both of these strings. But as long as you find the same order of the subsequence in you call it a longest common subsequence, in this case the length of this common subsequence turns out to be 4, which is also the length of the largest common subsequence. So, length of the longest common subsequence in this case is 4, ok. So, this is some random number if all these strings are chosen randomly, ok.

So, what is the length of a longest common subsequence and what can you say about the properties of such a random variable? Ok. Does it concentrate? Does it sort of deviate too much from let us say its central value, let us say its mean or it is median and if so then/what amount or how does it concentrate as a function of n maybe the alphabet size and so on, ok.

(Refer Slide Time: 30:02)



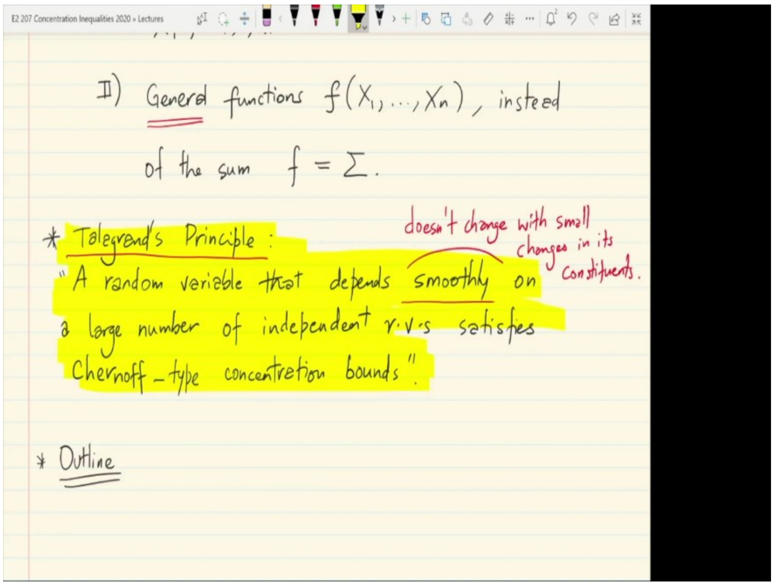
So, in fact, this course motivates the fact that to answer questions such as ah studying the fluctuations of the length of the longest common subsequence or let us say the number of coupons collected in entries or even the simple question of bounding averages or sums of random variables, we will find it useful to extend concentration of measure bounds to the following two situations.

One is when you have a quantity that depends on independently distributed random variables X 1 through X n, which need not be identically distributed. So, may all, they may all have different distributions. The more interesting in involved cases when you have to control the fluctuations of a general function f of X 1 through X n, where X 1 through X n are independent random variables, ok.

Instead of let us say the very simple sum that we have already studied which is f = the sum of all these random variables. So, you may have very complicated non-linear functions f of a bunch of random variables X 1 through X n and you may be wanted, you may be requiring to bound the fluctuations of f of X 1 through X n, which is just another random variable depending through f on these random variables.

In terms of you know some central quantity, some central measure like its expectation or its or its mean, ok. So, this is where, this is what the course hopefully will enable you to do and it will show you a spectrum of methods and techniques to try and break down the analysis of such the fluctuations of such functions.

(Refer Slide Time: 31:46)

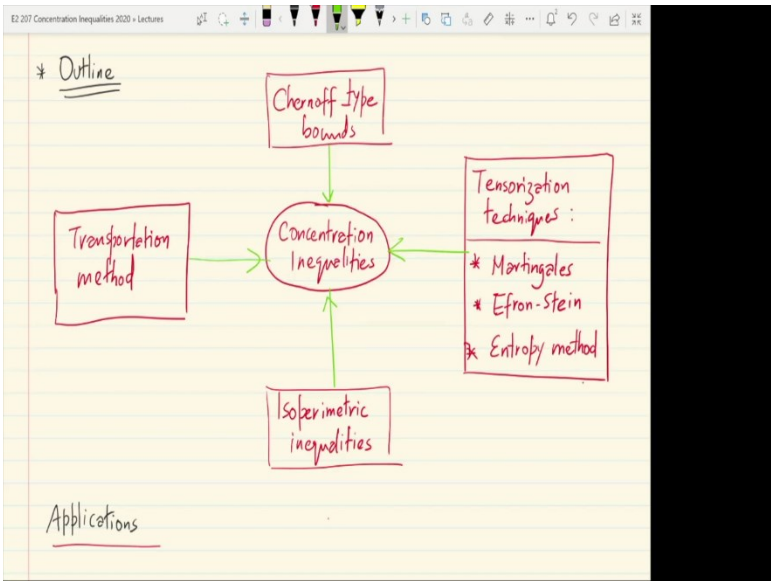


So, in this regard it is very important to keep in mind a very general principle that one of the great mathematicians of our era Michel Talagrand introduced. So, the insight that he obtained is that a random variable that depends smoothly on a large number of independent random variables satisfies Chernoff type concentration bounds, ok.

So,/smoothly, smoothly is used here as a vague term, but it can be made precise, but the point is that does it; smoothly here is used in the sense that its stable with respect to small perturbations of its constituent random variables, ok. So, does not change with small changes in its constituents, constituent random variables, ok.

So, this is the general principle that we will actually seek to elucidate and bring out for the entire duration of this course. And you will see this phenomenon happen repeatedly irrespective of whatever techniques we use to analyze the behaviour of a function of random variables that you know stability brings with it some kind of stability of the function overall brings with it a stability of the of its fluctuations.

(Refer Slide Time: 33:14)



So, just as a broad outline, I would like to present what kinds of a just a roadmap of what kinds of techniques one could expect to see in this course as it moves through the semester. So, at the simplest level we will have a simple Chernoff type concentration equalities that we will start out with, so Chernoff type bounds that we will build in some detail, ok.

So, this is one aspect of the concentration inequalities course. Moving on from here we will study broadly a technique called tensorization, ok. Tensorization techniques have to do with how you break down a complicated dependence of let us say one random variable on a bunch of others into simple manageable forms like sums, ok. The sum is probably the simplest way of breaking down an aggregate quantity into its constituents.

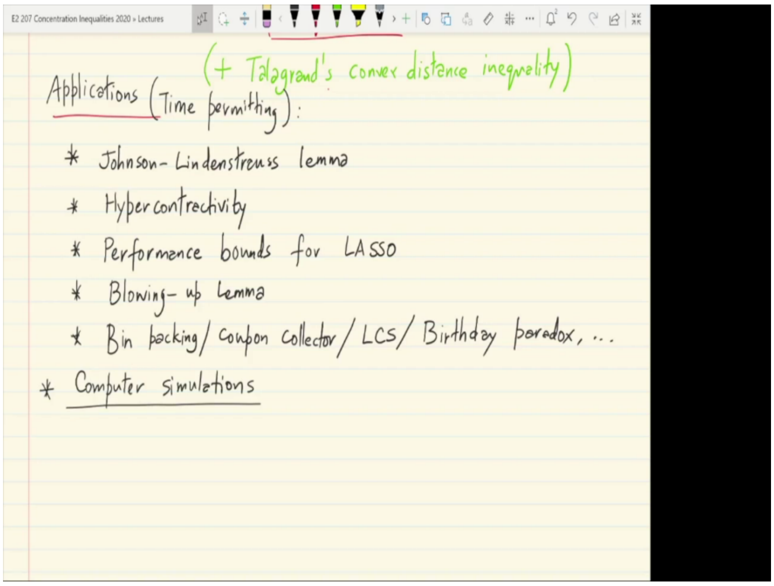
And we will study this in an abstract level in the form of what are called martingale bounds, then method called the Efron-Stein technique or the Efron-Stein inequality or family of inequalities and finally, conclude with what is called the entropy method, ok. So, this is going to be for a large part the second portion of the course.

Moving on we will also try to study the relation of what are called isoperimetric inequalities, which often arise in geometry and measure theory, how they connect to or how they can be useful in giving concentration inequalities. So, isoperimetric inequalities are an entire area that that generalize, that generalizes the you know common isoperimetric geometric inequalities that we have been used to since high school.

So, things like you know of all 2D shapes with a given area the one with the minimum circumference is the circle, the 2D circle, or conversely of all given 2D shapes with a given perimeter the one with the largest area is the circle. So, these are isoperimetric inequalities of a very simple form, but there are more general isoperimetric inequalities for measures over spaces that in fact, directly give us concentration equalities.

And we will also finally see what is called the transportation method, that relies on what are called transportation inequalities, which when used cleverly give you concentration inequalities to control fluctuations of random variables, ok. Along with this; so, this is in some sense going to be the core the core material of this course.

(Refer Slide Time: 36:37)



We will also hope to add time permitting things like a Talagrand's convex distance inequality which is a very celebrated inequality that in one sense tightens inequalities derived from the Chernoff method or martingale methods/using very different methods,/using a very different approach, ok.

Just a word about applications we will you know time permitting again, time permitting we hope to illustrate some of these concentration inequalities that we will uncover through the throughout the course in the form of applications to let us say physically relevant problems.

So, there are there is a long list of applications to which you can apply concentration inequalities. Some of these which we hope to touch upon are things like the Johnson-Lindenstrauss lemma, ok. So, this phenomenon is at the heart of many dimensionality reduction techniques in statistics and machine learning and big data something called hyper contractivity.

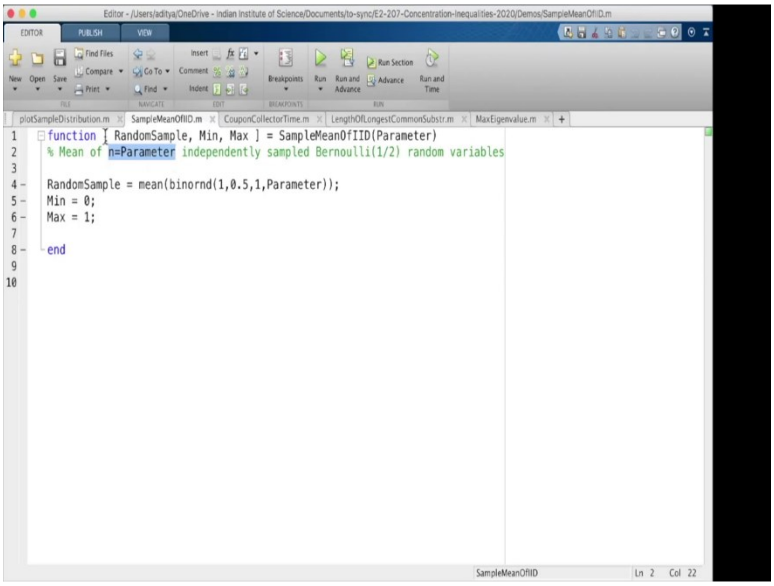
The analysis of the LASSO algorithm which is an estimation technique and its performance, turns out that some very non-trivial concentration inequalities help you understand its performance well, so LASSO. Then there is the application to deriving what is called a blowing up lemma or blowing up lemmas in which are useful in the context of information and coding theory.

Finally, there are also resource allocation problems in stochastic settings that benefit a lot from specialized concentration inequalities. So, these are let us say scheduling problems like bin packing, standard coupon collecting, longest common subsequence. Another name for coupon collecting is what is called the birthday paradox and so on and so forth, ok.

So, time permitting we will try to touch upon some or probably all of these applications. The last part of this lecture is something probably a more fun and hands on. I would like you, I would like to show you some actual computer simulations of random quantities like what we have seen before, like sums, averages the time to collect coupons, key coupons, even the longest common subsequence.

And show that show numerically that these quantities actually concentrate, ok. So, these quantities actually concentrate very tightly in practice and to some extent this course is trying to offer you the theoretical framework to try to make sense of these phenomena that are often observed in practical life.

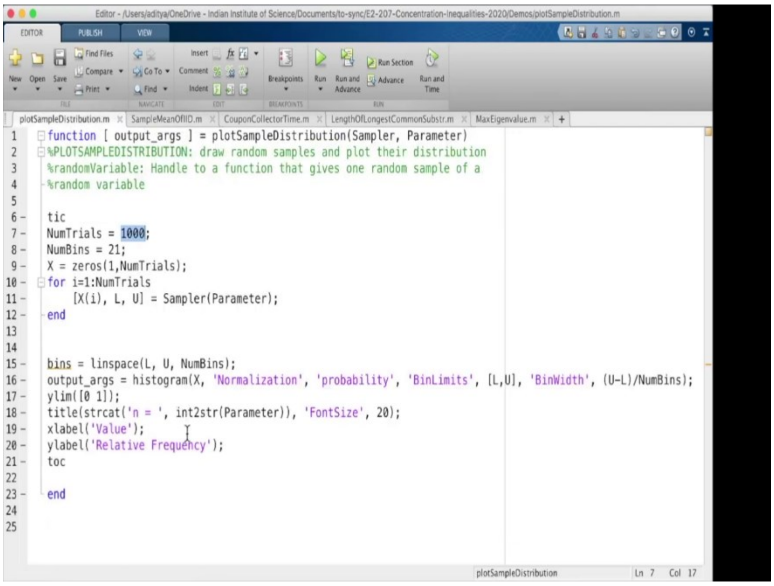
(Refer Slide Time: 40:03)



I am now going to show you some examples numerical examples of random variables that actually concentrate despite being complicated functions of other independent and simpler random variables. So, let us start in MATLAB/trying to show how the sample mean of a bunch of iid random variables concentrates.

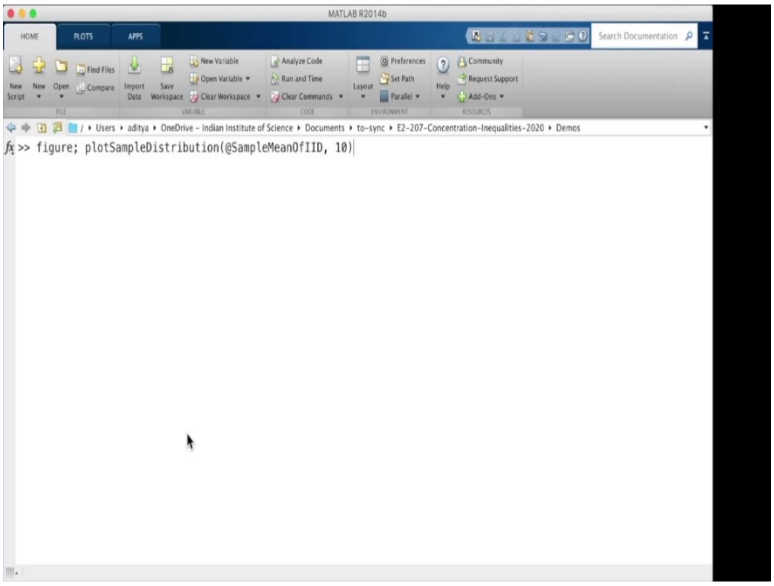
So, I have written this function here, which basically just takes every time it is called, it just samples n independent Bernoulli half random variables. Let us say representing the independent coin tosses that we saw in our first example today and just returns their mean, ok. So, this mean, the sample mean is also an independent random variable. So, each time this function is called it basically returns 1/n summation of X i, where the X i I have drawn iid Bernoulli half, ok.

(Refer Slide Time: 41:01)



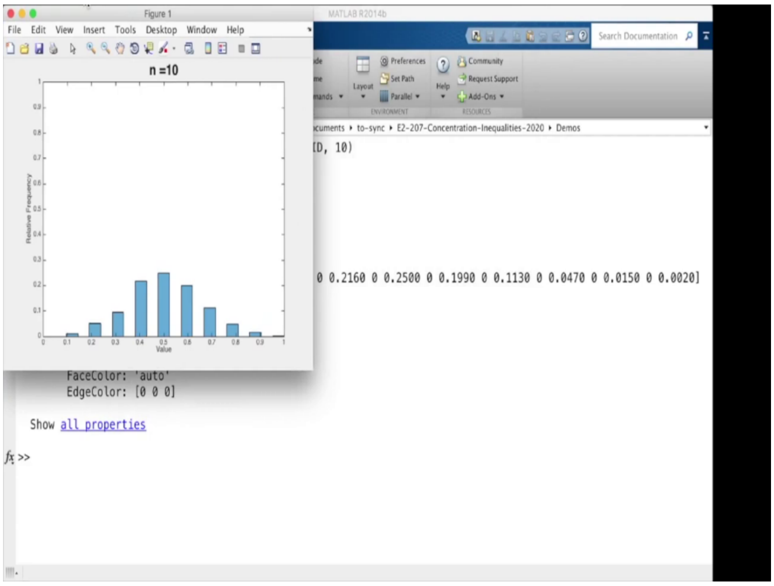
What I have written at the top level is a wrapper function called plot sample distribution that basically just takes 1000 samples independent samples of let us say this sample mean sampler, and then just plots a histogram of the samples returned, ok. So, it will plot essentially the empirical distribution or variation of all these 1000 sample means that are independently chosen with each experiment.

(Refer Slide Time: 41:30)



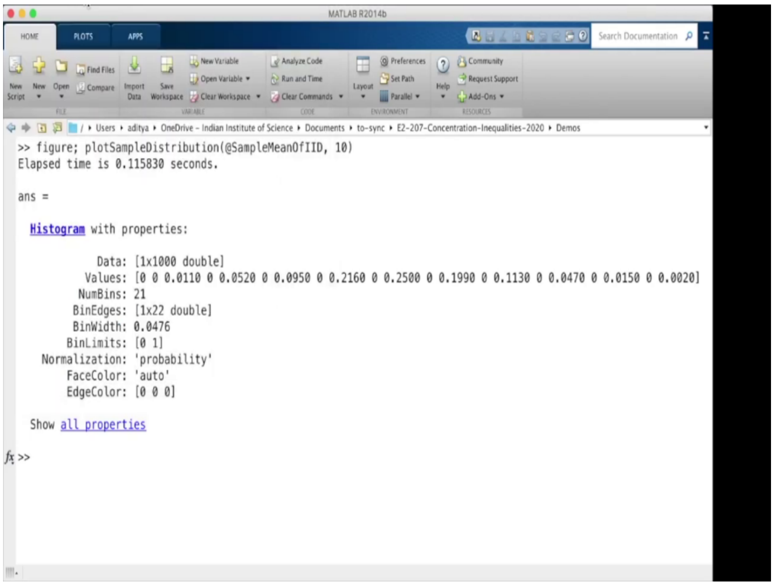
So, let us go a ahead, make a new figure, and plot the sample distribution for the sample mean of iid random variables, where n is set to be let us say 10, ok. So, 10 random variables are going to be sampled and their means are going to be plotted for 1000 separate experiments, ok. So, this is what happens, ok.

(Refer Slide Time: 41:56)



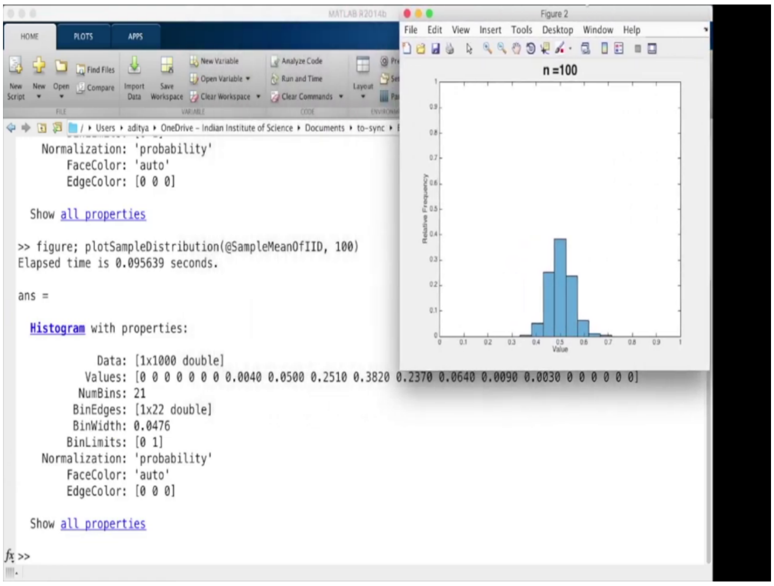
So, you see this histogram that is basically between the values around the X axis between 0 and 1 and as expected the peak is around 0.5, ok. So, a large number of the sample means fall near in the neighbourhood of 0.5.

(Refer Slide Time: 42:19)



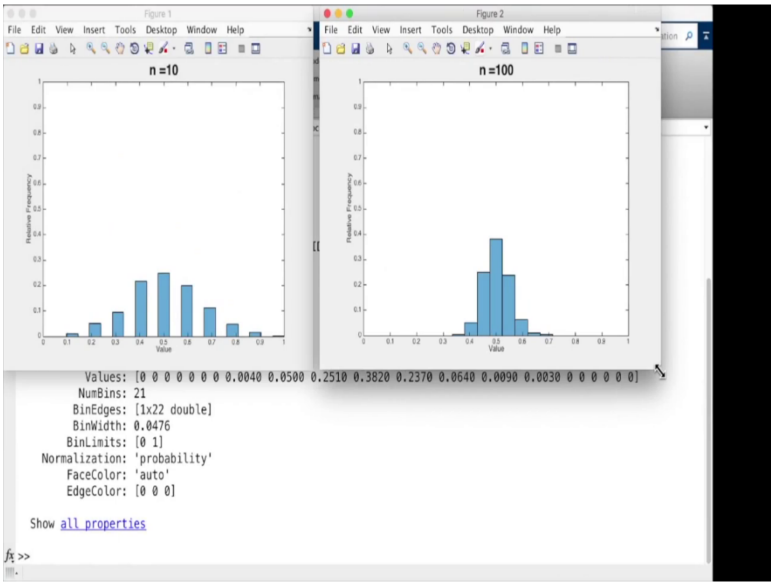
Let us now increase the number of coin tosses from n = 10 to let us say n = 100 and then see what happens.

(Refer Slide Time: 42:29)



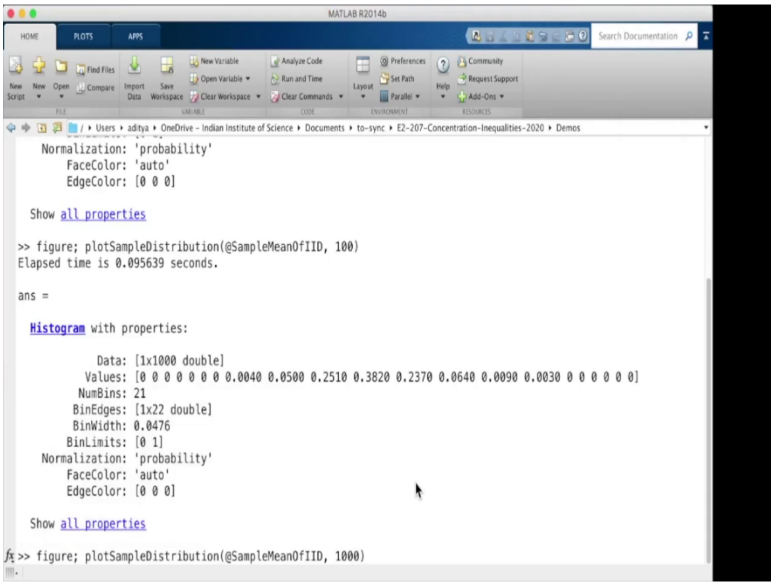
So, I am just going to increment the number of samples taken to compute each sample mean from 10 to 100, ok. And so, this is what happens. You can see that it is again distributed close to 0.5, but then much better concentrated.

(Refer Slide Time: 42:41)



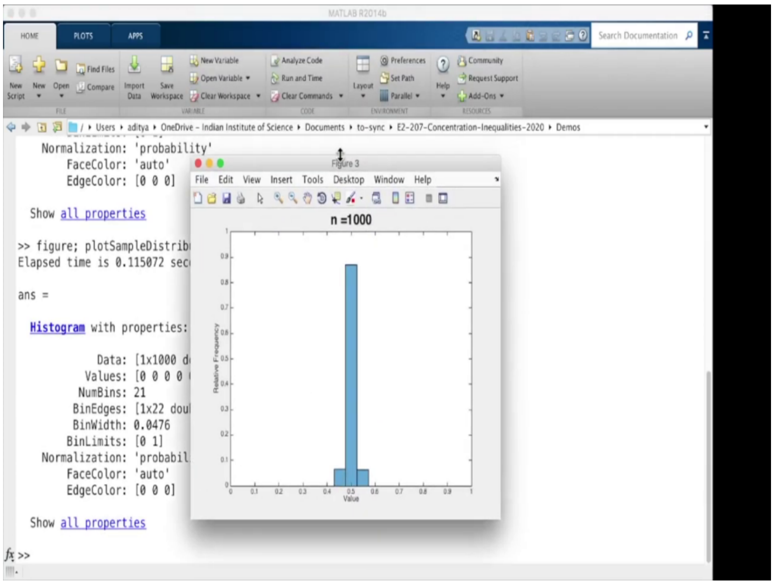
So, if you see them see these plots side/side, the first one and the second one, you will see what I mean here, ok. So, going from 10 coin tosses to 100 coin tosses has already made the sample mean concentrate somewhat.

(Refer Slide Time: 42:52)



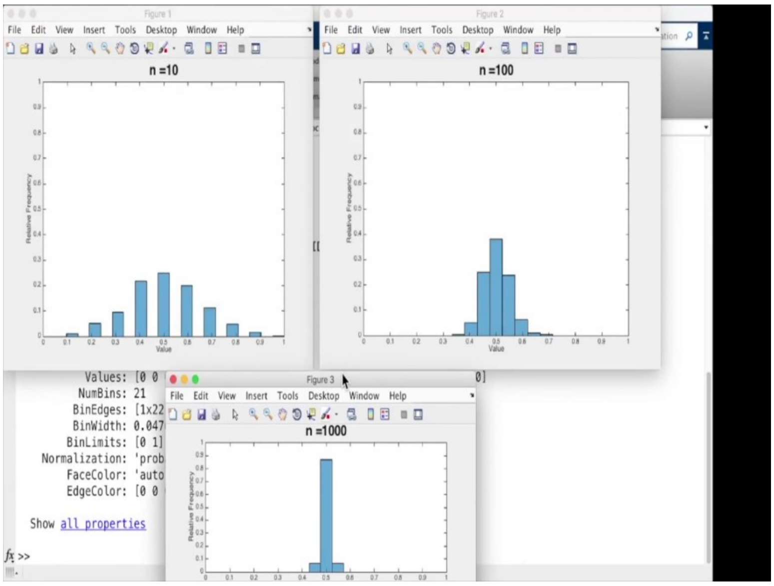
Let us do one last experiment where we will take 1000 coin tosses and plot sample means for these 1000 coin tosses in 1000 distinct independent experiments.

(Refer Slide Time: 43:05)



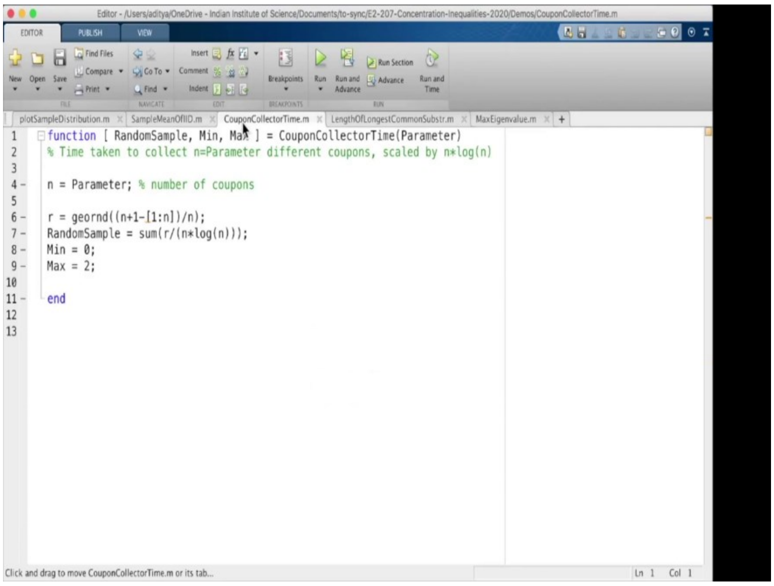
So, this is what happens. You can see that there is a very sharp concentration.

(Refer Slide Time: 43:15)



And if I just get all these figures together you can see that this concentration of measure phenomenon is actually happening. So, there is most of the probability very very close to 0.5, if you toss more and more number of coins.

(Refer Slide Time: 43:40)

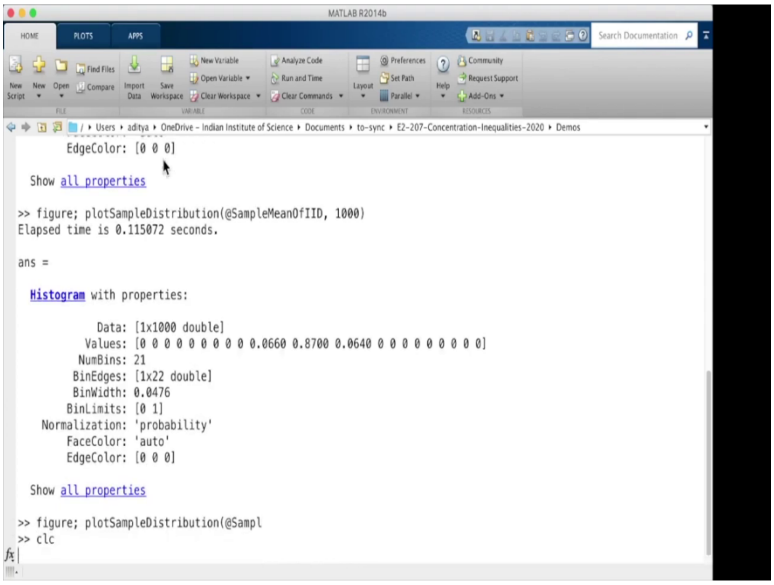


So, let us move on to a slightly more complicated example, which is also what we saw in the lecture and this is basically going to be the time taken to collect; this is a random variable representing the time taken to collect n different coupons, ok. So, each time this sub function which is coupon collector time is called what it does is it simulates a random experiment where there are n different coupons, ok.

You can think of these as n different colours of coupons. In each time it basically collects an uniformly random coupon labelled anywhere from 1 through n, and it basically measures the first time at which you collect all n coupons, ok. So, this is the time at which you are happy to have collected all the colours.

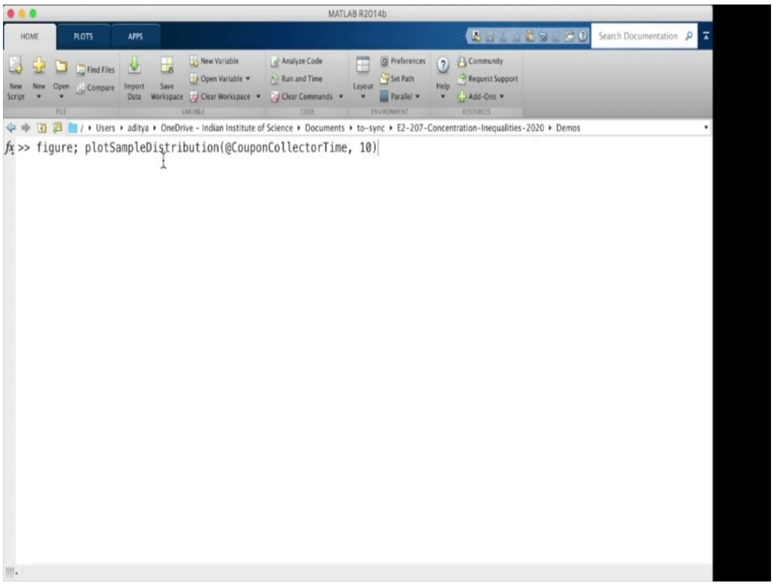
And for technical reasons we are going to scale this first time at which you collect all n coupons/the number n log n, ok. n log n in some sense is used because we know that the expected time to collect all the n expected first time at which you collect all the n coupons is of order n log n, ok. So, it is reasonable to scale/this factor, ok.

(Refer Slide Time: 44:47)



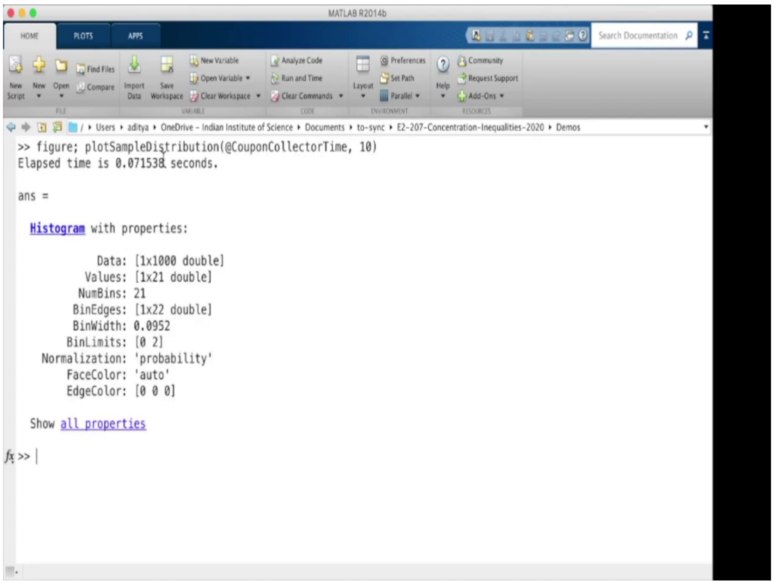
So, let us go ahead and see what happens when.

(Refer Slide Time: 44:55)



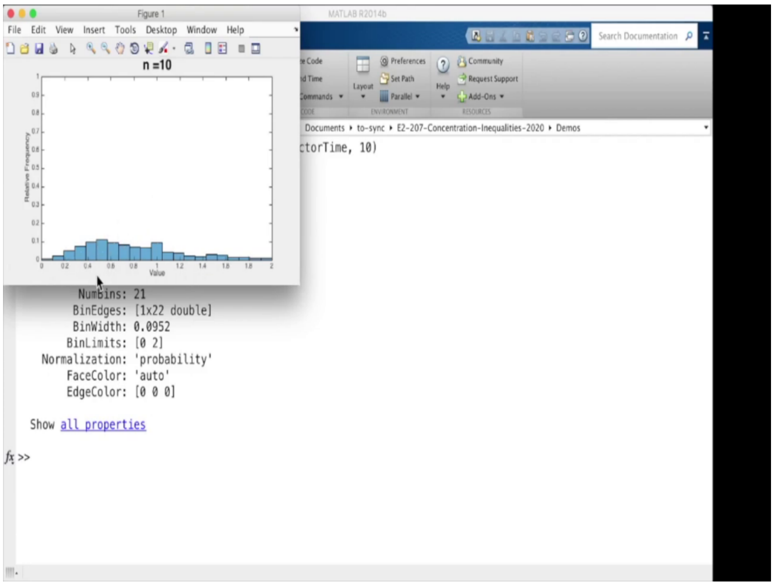
So, I am going to change, I am going to change to plotting the coupon collector time; when let us say the first time at which you collect coupons of colours between 1 to 10, ok.

(Refer Slide Time: 45:03)



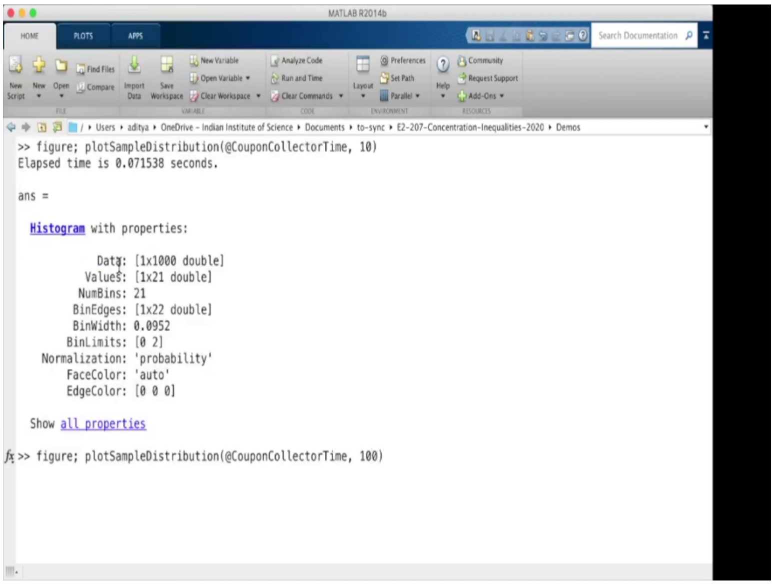
And so, this is what happens. You basically collect coupons at times roughly spread out from 0 to 2 times n log n, where n is 10, ok. So, roughly 2 to 20.

(Refer Slide Time: 45:10)



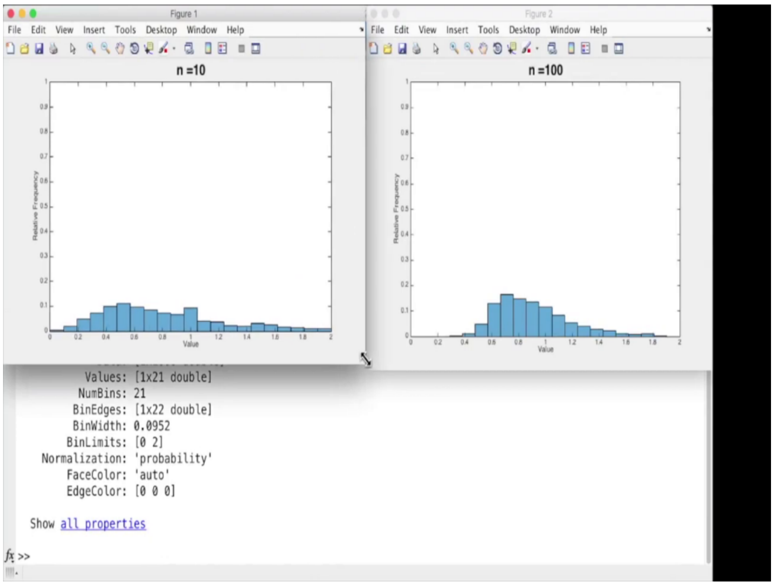
And this is the probability distribution of how likely it is that you collect the coupon at a given time of X axis.

(Refer Slide Time: 45:34)



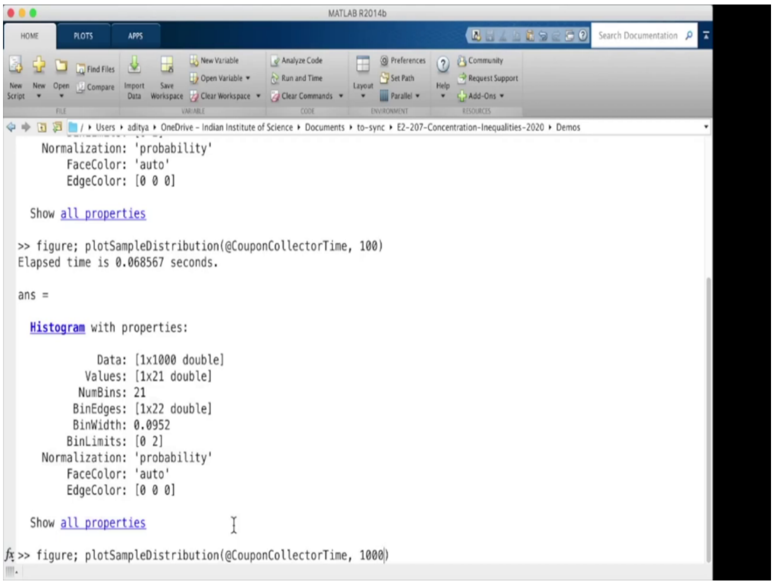
So, let us bump it up to collecting coupons of 100 different types.

(Refer Slide Time: 45:40)



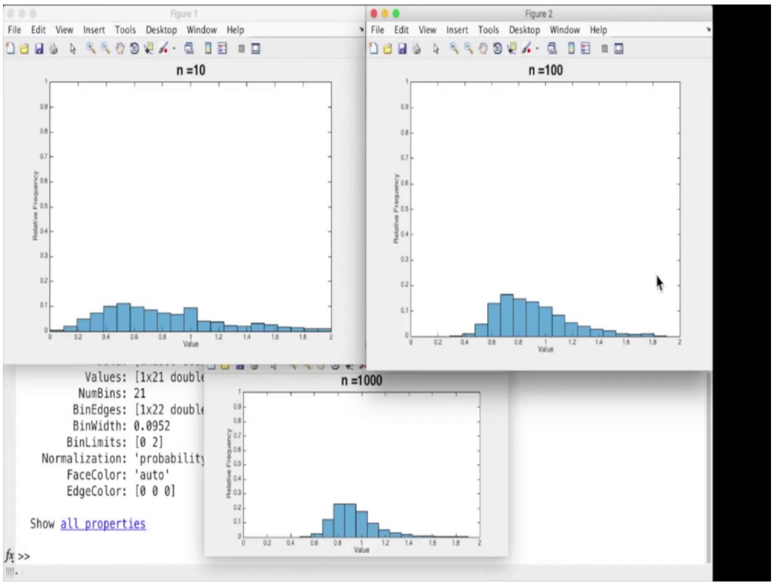
And so, this is what happens. You can see that it started to concentrate a little bit. If you see this in context, this is evident, ok.

(Refer Slide Time: 45:55)



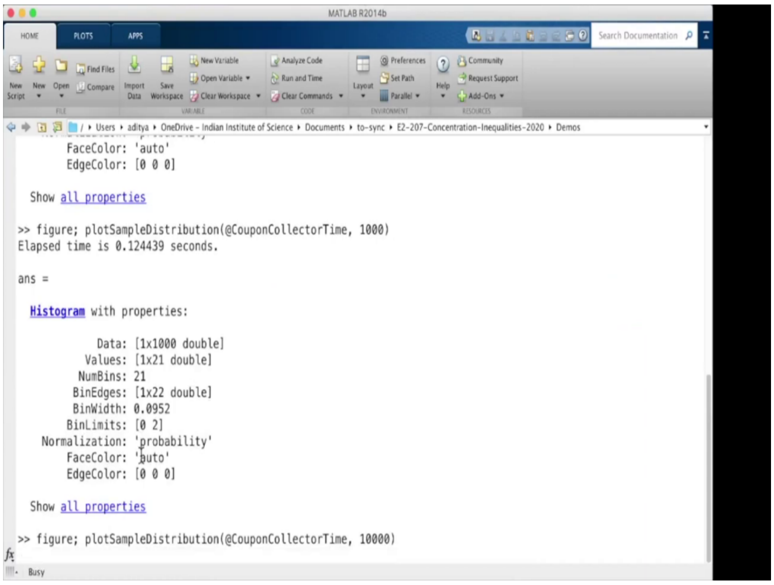
Let us maybe see what happens when you try to collect; you look for the first time the random time at, which you collect 1000 different coupons, ok.

(Refer Slide Time: 46:00)



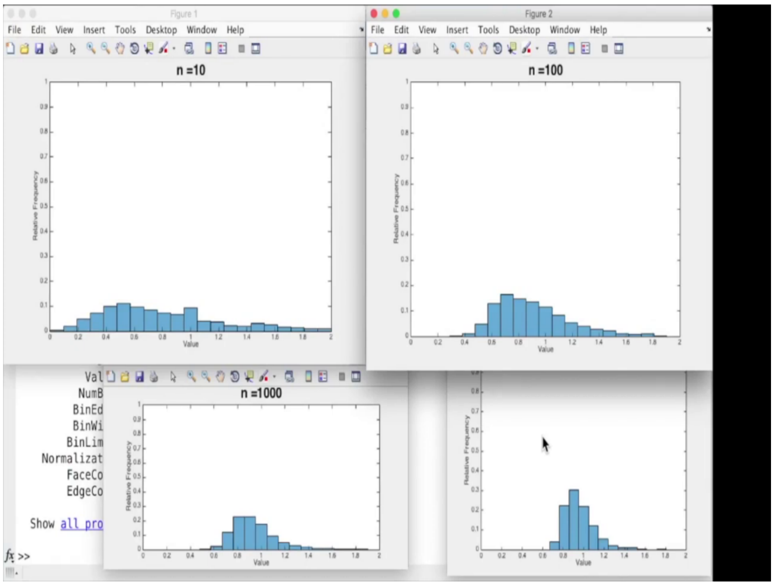
And so, you can see that this is already concentrated, ok. And so, that is what is happening, ok.

(Refer Slide Time: 46:15)



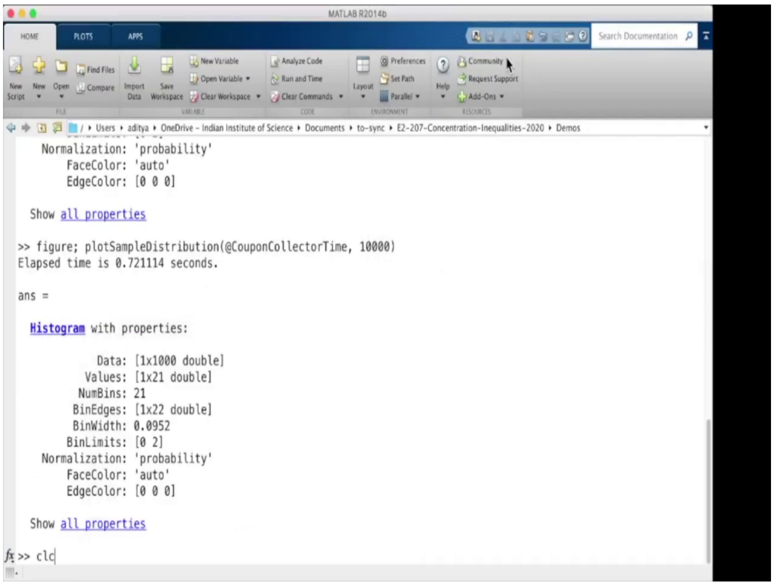
Maybe let us do one more while we are at it.

(Refer Slide Time: 46:18)



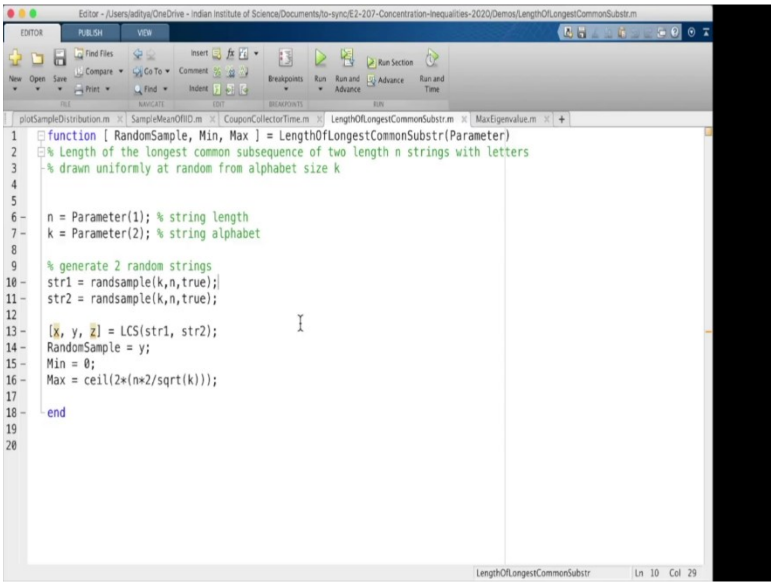
So, 10000 coupons, just to show you the sequential progression, ok. So, on the top right is 10, then there is 100, and there is 1000 and there is 10000, ok. So, you can see that there is a concentration of probability phenomenon happening and you can actually use the techniques that we will go through in the course, some of these techniques to try and actually get a theoretical explanation for this phenomena, ok.

(Refer Slide Time: 46:56)



So, let us move on to our third numerical example, which is also something we saw in the lecture, which is the problem of the length of the longest common subsequence, ok.

(Refer Slide Time: 47:04)

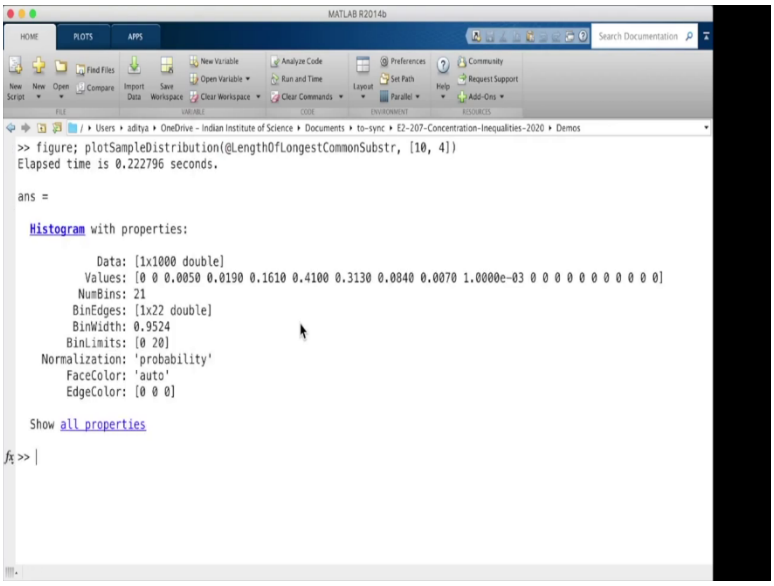


So, what is happening here is that each time this experiment is run, what it does is the following. It takes two strings of a specified length n where each element of each string is drawn uniformly at random from an alphabet of size k, ok. So, if k is 2, each is a binary string if k is = 3, then each of these n length strings is a ternary string and so on, ok.

So, generate two random strings and then basically do an operation to compute the length of a longest common subsequence between these two length in strings. And so, that is what it is going to return, ok. So, we are going to directly histogram or plot the empirical distribution of 1000 independent experiments of finding the length of the longest common substring or subsequence, ok.

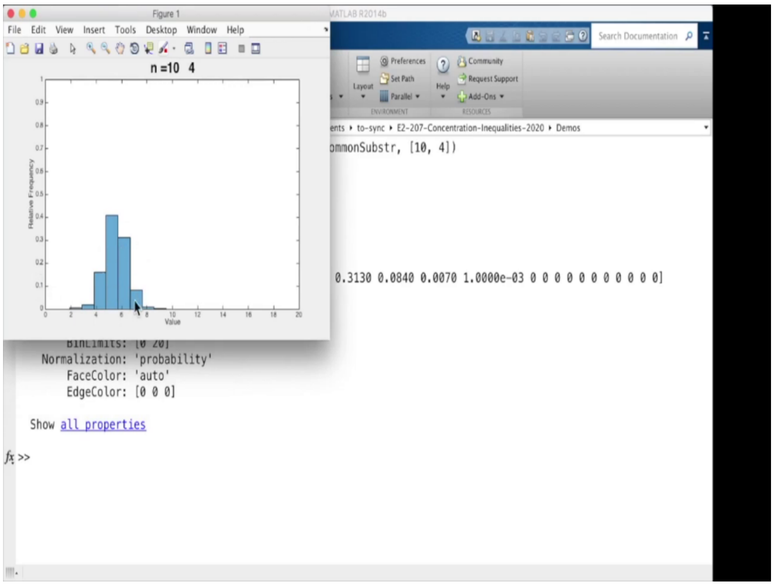
So, recall that the subsequence need not occur contiguously in both of these strings that is/the definition of the longest common sequence problem.

(Refer Slide Time: 48:11)



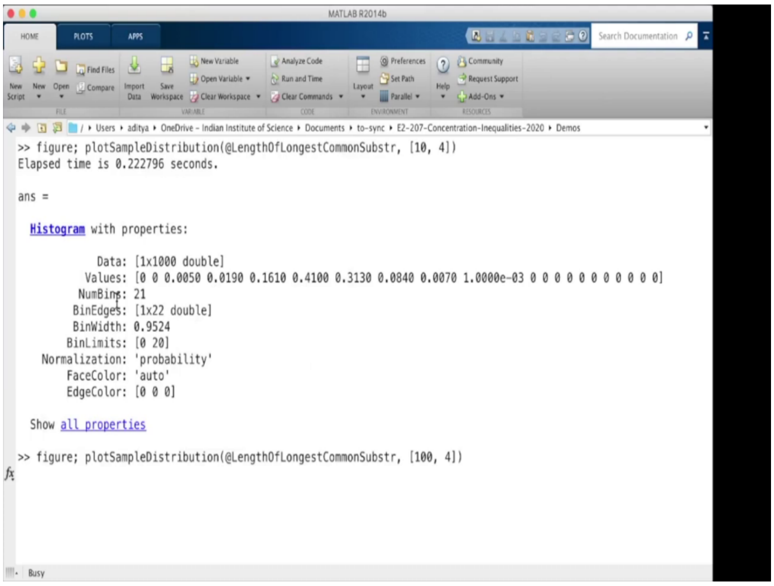
So, let us go ahead let us go ahead and do the same histogramming exercise for the length of the longest common substring. Let us say we generate strings of length 10. Let us say with quaternary alphabet. So, each symbol can be 4 letters, and these are n length strings, ok. So, 4 letters see each can each symbol can either be 1 of 4 letters, just as let us say you have in the DNA sequence or the RNA sequence.

(Refer Slide Time: 48:54)



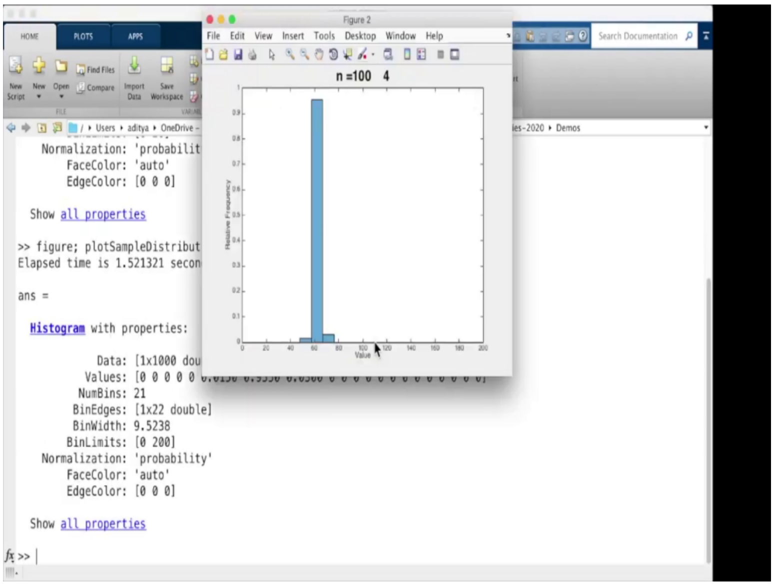
So, this is the histogram for the longest common subsequence between two length 10 strings, ok. It is obviously, a number between 0 and 10, and it is concentrated somewhere somewhat concentrated somewhere around 5 or 6.

(Refer Slide Time: 49:12)

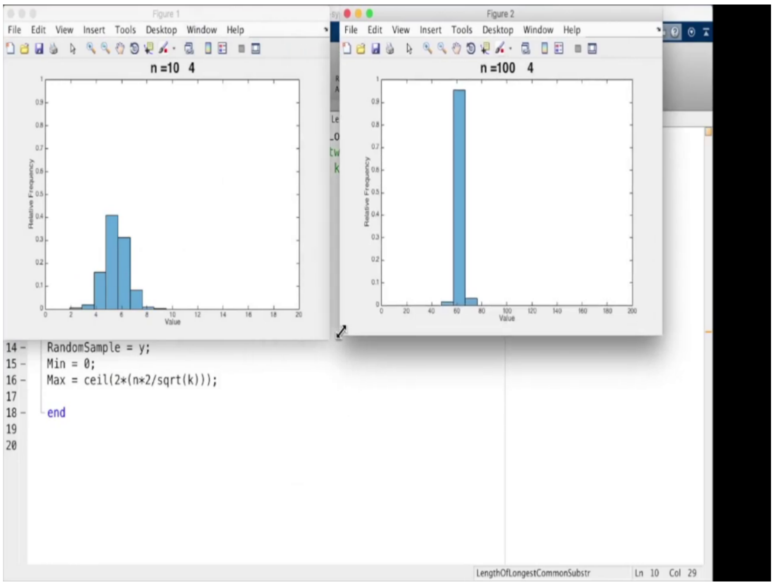


Let us bump this up to finding the length of the longest common substring between to length 100 strings, ok.

(Refer Slide Time: 49:25)

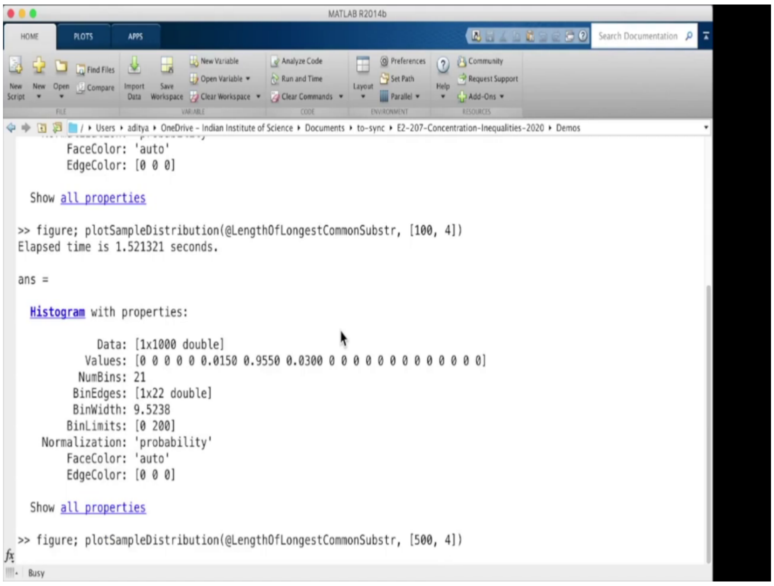


(Refer Slide Time: 49:19)



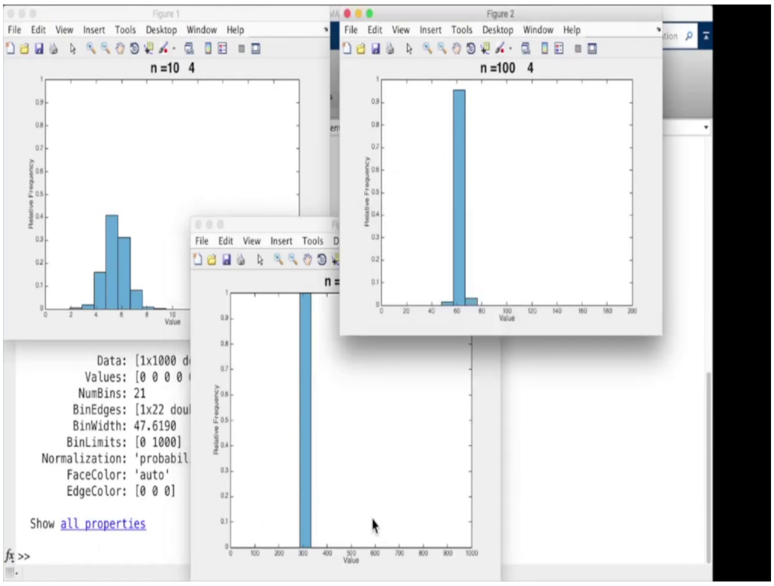
And you can already see that there is a sharp concentration, so let me put these side/side, so that you can compare these plots, ok. So, there is some fairly sharp concentration occurring at around length 60 years or 65.

(Refer Slide Time: 49:44)



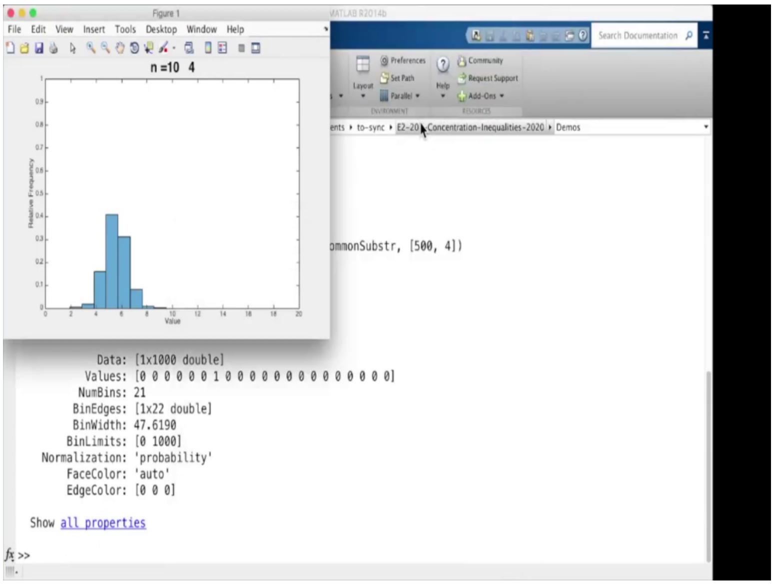
I do not know if this will work for length =; let us try length = let us say 500 assuming it does not take too long. So, we need to let the sampling run for some time, ok. There it is much more concentrated, ok.

(Refer Slide Time: 50:00)



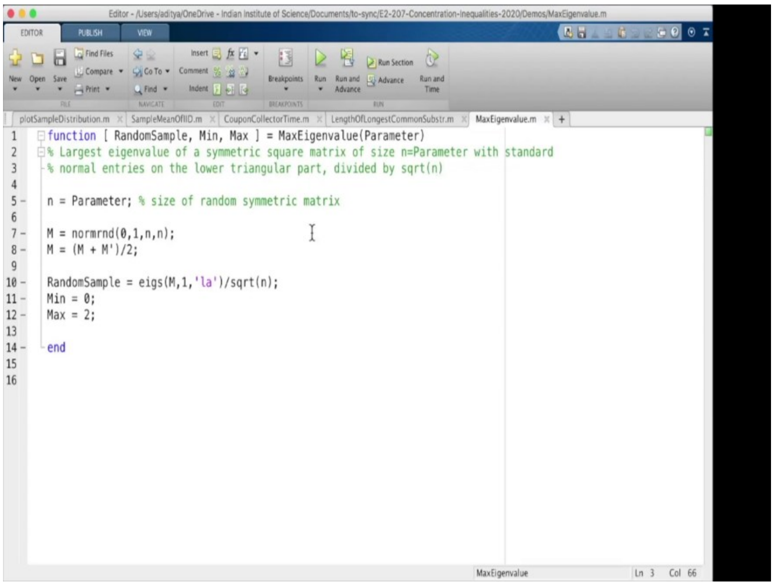
So, you can see that there is barely a difference going from 100 to 500, but nevertheless all the probability has basically moved into the central peak, ok.

(Refer Slide Time: 50:18)



If you have a more powerful computer maybe you could try more fancier experiments.

(Refer Slide Time: 50:28)

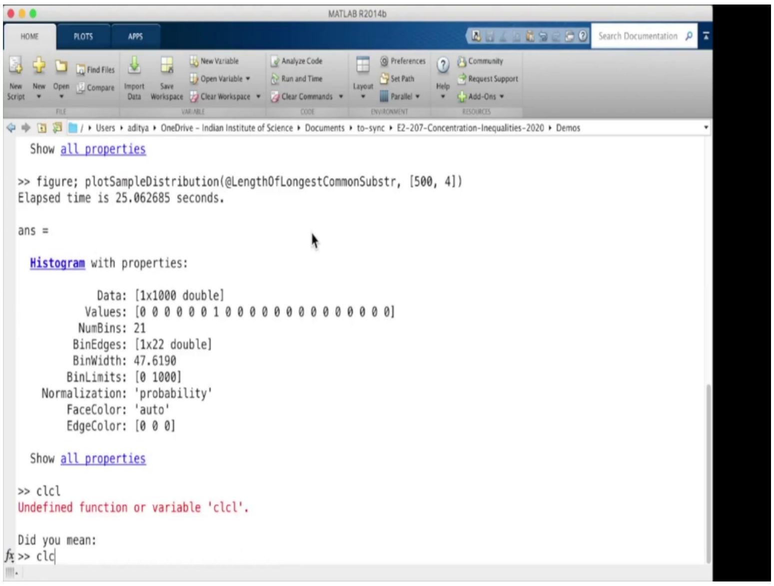


Let us see. I had one more here to present which is again a very interesting random quantity that occurs in all aspects of engineering and random structures which is the largest eigenvalue of random matrix of random matrix, ok. So, here is a specific example where the specific experiment run each time is to return a random variable which is the largest eigenvalue of a symmetrically generate, a symmetric randomly generated square matrix of size n.

So, it is an n/n symmetric square matrix which is generated as follows. So, the lower triangle is populated independently with standard normal entries, ok, and those entries are basically mirrored on the upper triangular part. So, that you get a symmetric random matrix, ok.

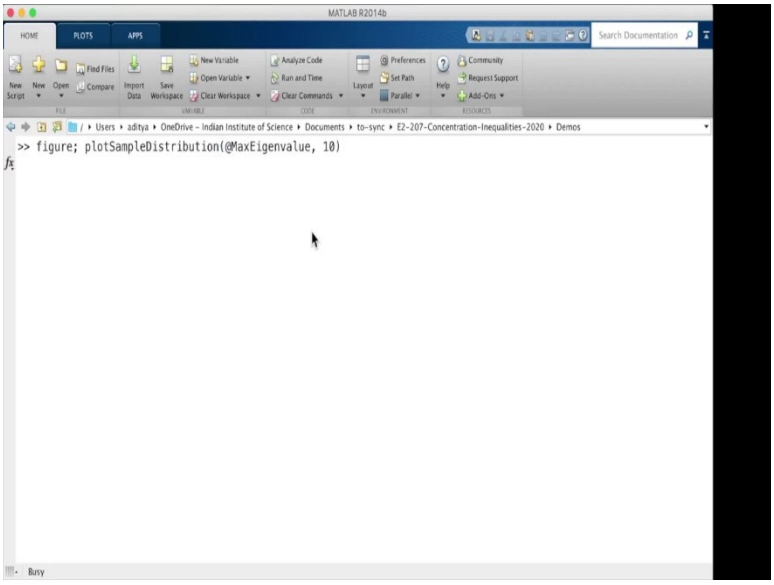
And for technical reasons we are also going to divide this largest eigenvalue of this n/n matrix/the quantity √ n and then plot the histogram of this random quantity, ok, the largest eigenvalue of a random matrix, ok. So, in this case this is also connected to the norm. Largest eigenvalue of this matrix is a measure of its size, ok. It is a single scalar measure of how large this matrix is this random matrix is, ok. And we would like to study the distribution of this quantity and whether it concentrates well.

(Refer Slide Time: 51:50)



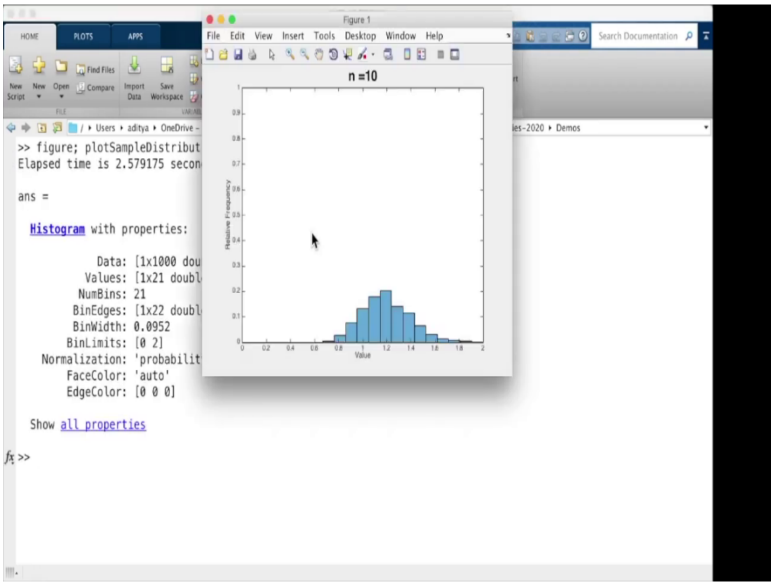
So, let us go ahead and run our sampler and histogramming routine, wrapper routine for the eigenvalue example.

(Refer Slide Time: 51:54)



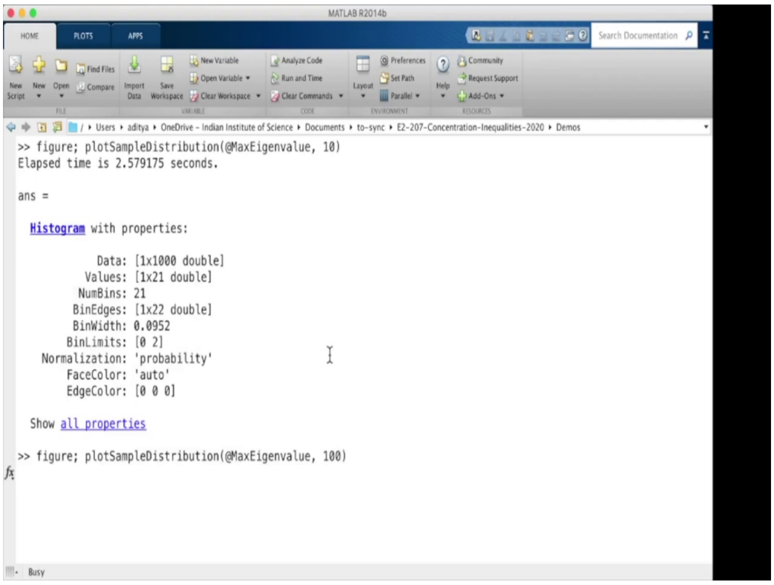
So, we will pass it a matrix size of let us say 10. So, this is what happens.

(Refer Slide Time: 52:12)



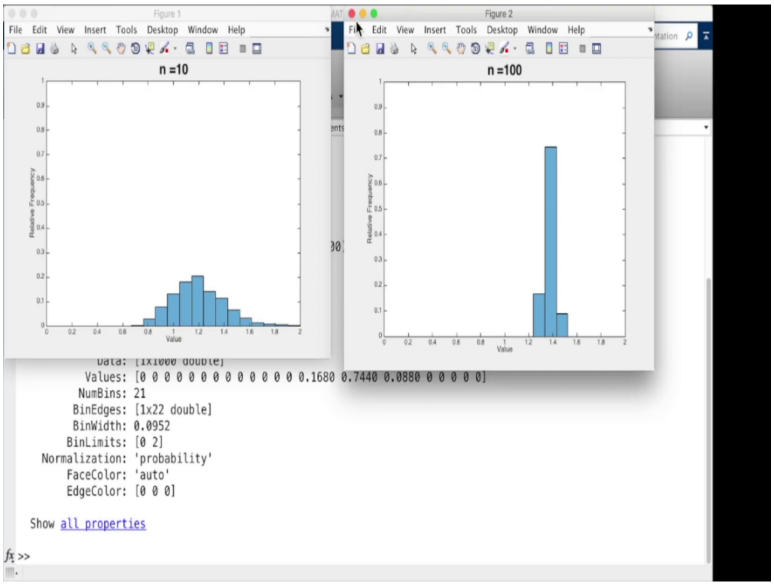
So, for a 10/10 matrix generated randomly according to the description I gave you, this is what happens to the normalized maximum eigenvalue, normalized base √ n.

(Refer Slide Time: 52:29)



If we run this for matrices of size 100, we need to give it some time to compute and sample, ok.

(Refer Slide Time: 52:46)



So, it is already looking like its concentrating fast, ok. So, we can see that there is a distinct pushing towards of all the probability mass towards sort of central value and this is going to get even more concentrated in (Refer Time: 53:09), the size of the square matrix increases, ok.

So, there are many such examples. And hopefully the techniques that we will study in this course on concentration inequalities will give you a basis to reason about why these phenomena occur and when you should expect such concentration phenomenon to occur in problem setting with this study.

Thank you.