Introduction to Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department of Computer Science and Engineering

Indian Institute of Technology, Madras

Module - 02

Lecture - 08

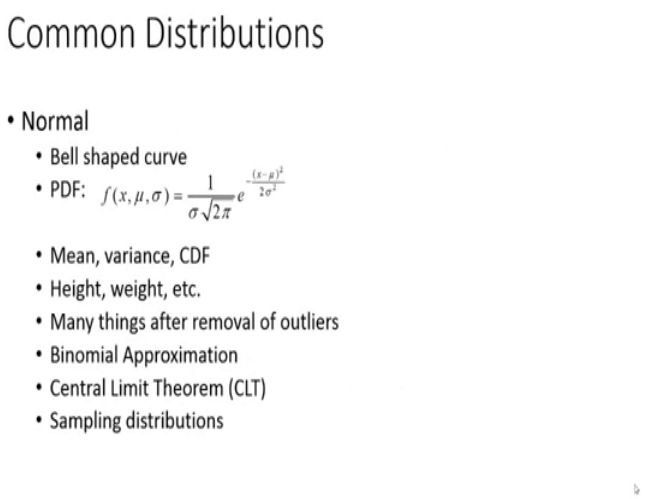
Random Variables and Probability Distributions-3

Hello and welcome to the third and last lecture on the series on Random Variables and Probability Distributions. In the first lecture we spoke about, we introduced the concept of random variables spoke about, how probability distribution can be discrete or continuous and we also introduced the idea of PDFs and CDFs, Probability Density Functions and Cumulative Density Functions. In the second lecture, we targeted about five or six distributions, commonly used distributions and we introduced them.

As well as talking a little bit about, how one can get the CDF if you are given the PDF, what is the relationship between the PDF and CDF and vise versa and how do you get to the PDF given a CDF, symbolically. And we also spoke about, how you can mathematically using given a distribution compute it is mean, compute it is variance and so on. In this lecture, we are going to focus more on a single distribution called the normal distribution, many of you might have already heard about it.

But, we are also going to look at some applications associated with this distribution and one really important application has to do with inferential statistics, which is something that will be quite central to the next 4 or 5 lectures. So, it is in that idea that, we are introducing the normal distributions.

(Refer Slide Time: 01:36)



So, the normal distribution itself you might have come across it, if you not you might have heard of this thing called the bell shaped curve. So, the distribution itself looks like the shape of a bell. So, just like the uniform looks like a flat line and you know different distribution have different shapes, this looks like a symmetric bell, bell shaped curve and the probability density function of this distribution is characterized by this formula. This formula that is shown here and one thing that is noteworthy is that, this distribution has two parameters mu and sigma.

So, the distribution itself is defined by the mean and variance, so the mean and variance of this distribution go into the formula and they defined it. So, there is no point saying tell me, what is the probability of value x for a normal distribution, because that question does not make sense. In order to say for a distribution with this mean with this variance, what is the probability of value equal to a greater than x?

So, that question means more or you know, what is the probability of finding a value between x and x plus delta, for a normal distribution with a mean mu and a sigma equal to sigma and standard deviation equal to sigma. But, once you given the mean and sigma, it is quite simply this formula that you would use and you can compute the probabilities. So, what is the mean of a particular normal distribution defined by mu and sigma?

Well, that is very straight forward, it is the mu, because the distribution is defined by mu and the variance is nothing but, your sigma square. So, you can, it is quite straight

forward there is well.

The CDF; however, is not something that simplifies very elegantly. So, to define the CDF you would still use your traditional procedure of using the integral and by the way the normal distribution goes from minus infinity to plus infinity, so it make sense to actually use the minus infinity here. So, you would actually use the minus infinity to x, f of x, which is the PDF, which is nothing but, this formula, so dot d x.

But, while in many distributions this whole thing simplifies and you are able to do the integration and there is an actual value, with the normal distribution it does not simplify very elegantly without using more complex algebraic terminology. So, the CDF is often just stored in tables, sometimes especially for the normal with mean 0, standard deviation 1 or it is just something that you integrate each time to get.

Now, this is a very interesting distribution, because there are lot of things that are normally distributed. So, things like peoples height, weight well height; obviously, with each gender, grades in a class, marks that people score in exams. The core idea with the normal distribution is that, unlike the uniform distribution, which says everything is equally likely.

It is the normal distribution says that things in the extremes are less likely, things in the center are more likely within certain limits, which is what gives it it is characteristics bell shaped curve. I mean, if this is any attempt to the bell shaped curve, we basically saying that things that are on the extremes, like here and here are less likely and things in the center like here are more likely that is why they have a greater height, with all of these things the y axis is the probability.

So, if you take a look at something like heights or let us say weights and you fix a gender, let us say male and you take something like people, who are registered for introduction to data analytic course, then you will find that there might be very few people, who weigh less than I do not know 40 kg, so or 50 kgs, men especially. And you will find very few of them probably weighing more than 100 kgs or so and then, you know and so that kind of tapers off, an either extreme you find less, in the center you find more.

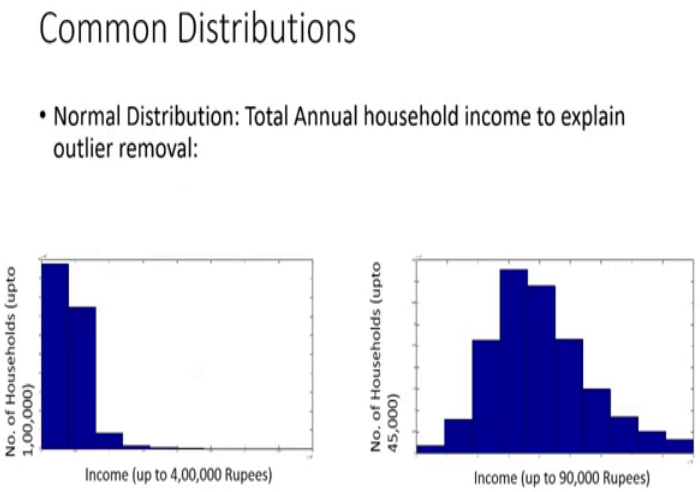
But, there are many other distributions are also like this, but this that this is that is key feature of being in a bell shaped curve. The other thing is you know many things after

you remove outliers start to look normal and we will talk about an example of that. In this slide, I am just not going to talk about the other things that we will talk about in this lecture, so I am not kind of rushing through it. We will especially take up from here and till here and go through them in detail with slides.

But, you are also encounter that there is this things called the binomial approximation, which is… We briefly spoke about this when we introduce the binomial distribution that certain problems, which just by definition look like they fall so cleanly as a binomial distribution, for computational reasons could be quite easily approximated to a normal distribution. Although, the binomial is a discrete distribution and the normal is a continuous distribution.

We will also talk about something called the central limit theorem, which makes the normal distribution very useful for many applications and also a very interesting concept per se and finally, we will look at the idea of sampling distributions. The core idea being that, if you take a random sample of size x of associated with any variables, so I randomly select five people and measure their heights. Is there a distribution associated with the parameters that I get like the mean and standard deviation? But, we will talk about this in greater detail.

(Refer Slide Time: 08:04)



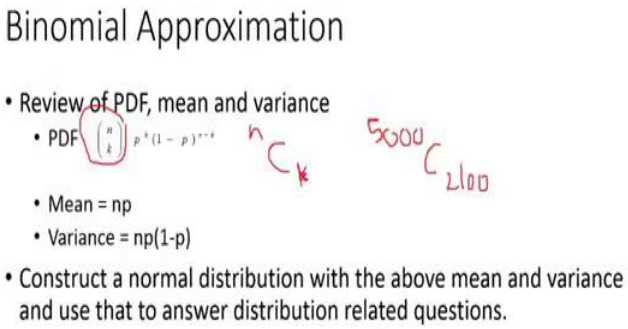
So, the first thing is things after removal of outliers. So, here is an example of some real data, where we looked at the total annual household income and you know, so the graph

that you see to the left hand side is you know, it is essentially all these households with income up to and we just stop the x axis at a certain point and so, we said let us look it income up to a certain value and the y axis is the number of households.

So, I have created essentially a histogram, but that is a proxy for finding the probability distribution itself. So, you can think of the probability distribution as something that looks like this, in this particular case. People cannot have incomes less than 0, so on and so forth. Now, look at the same graph, where I said I am not going to look up to 4 lakh rupees income, but I am just going to concatenate the x axis in 90000 rupees.

So, the whole idea was to say that some of these values could have been outliers and we took a certain value beyond, which we go. And already you can see that this graph is starting to look a lot more bell shape ((Refer Time: 09:22)) Probably not perfect, but the core idea is this, which is that sometimes once even though the distribution originally might not look normal with sufficient amount of outlier removal, the distribution could truly be know.

(Refer Slide Time: 09:38)



The second concept that we want to speak with respect to this is the binomial approximation. So, let us just very quickly review, what the binomial distribution is about. We spoke about, how this term in the PDF of the binomial distribution was really, n choose k. So, n combinations, k combinations out of n was the core idea and that is fine. So, if you have problem of the type saying, what is the probability of finding, you know

3 heads out of 10 tosses. This works fine, you can substitute the values get the PDF.

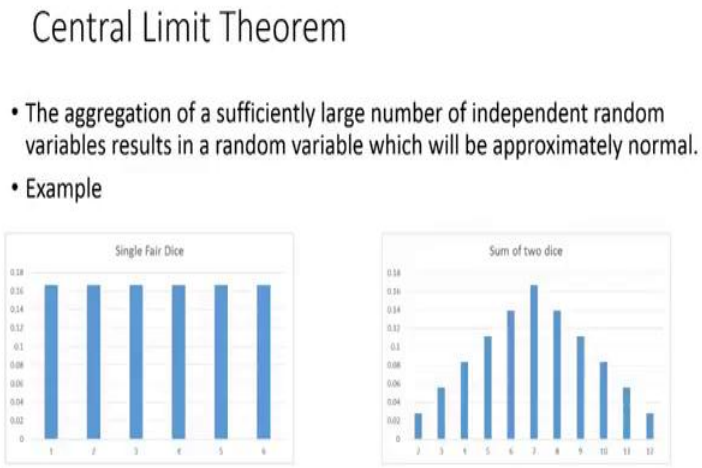
Now, somebody came and asked you saying, what is the probability of getting 2100 heads out of 5000 tosses. Then, you essentially need to, if you want to use this formula you need to plug in 5000, you know c 2100 or whatever the number is and you know; that is a very large number; that is a very hard computation and you could 5000 and 2100 just an example that could be 5 million and you know 200000 and it is very hard to do those calculations.

So, one thing that you can do, when n becomes really large is you can essentially use this formula that you have for mean and variance of the binomial distribution and construct a normal distribution with this mean and this variance and used to answer distribution related question. So, you for instance if there is a 50 percent chance for instance of a coin falling head and tails, you can say well the mean of 5000 tosses is 2500, because you have 5000 tosses times 50 percent probability. So, that is 2500 and that is your mean and your variance also you would similarly calculate by plugging in n equals 5000 and p equals 0.5.

And once you do that, you can essentially construct a normal distribution with these parameters and you can answer questions like, what is the probability of there being more than 2100 heads or what is the probability that the number of heads would be between 2000 and 2500 out of 5000 tosses. You; obviously, cannot answer a question like, what is the exact probability of getting 2112 heads, because you essentially converted this to continuous distribution.

And the idea of answering a question like, what is the exact probability of 2121 tosses out of 5000 or I mean 2121 heads out of 5000 tosses becomes relatively meaningless, because as n keeps becoming large the probability of any one thing exactly occurring becomes really small becoming close to 0. So, you are interested more in intervals, which is in spirit this, what you can do with continuous distributions and you can use a normal approximation of the binomial distribution to achieve that as long as n is fairly large.

(Refer Slide Time: 13:00)



Next, we will move to something called the central limit theorem and the core idea here is that the aggregation of a sufficiently large number of independent random variables results in a random variable, which will be approximately normal. So, what is that mean? It just means that look, if you have some process and it is some distribution from that process, so let us say flipping a coin or throwing a dice is the process. Now, central limit theorem says as long as I am aggregating many such processes.

So, if I said instead of asking you the simple question of the distribution associated with what I would get, if the roll the dice once. I instead say I want to know the distribution associated with rolling the dice twice and I am going to add them up. So, the first time I will roll the dice and then, I get some number I write it down, I will roll the dice another time and I will get another number and I am going to add those two numbers.

Now, the distribution associated with that sum is also probability distribution, because you know it is still a random process; there is still some chance that I can get each value. I clearly cannot get any value less than 2, because first time I can roll 1, second time I can roll 1. So, I cannot get 1, I can only get 2 as the minimum value and the maximum value is 12, I can roll 6 and 6 and that is 12.

So, the idea that is being put forth here with central limit theorem is that aggregating it and the word aggregating can be thought of is, you know taking the sum or you can think of it is taking the average, both a forms have, both are essentially the same thing. The

difference between sum and average is, average is just divided by the number of times. But, this form of aggregation of a sufficiently large number of random variables results in a random variable, which will be approximately normal.

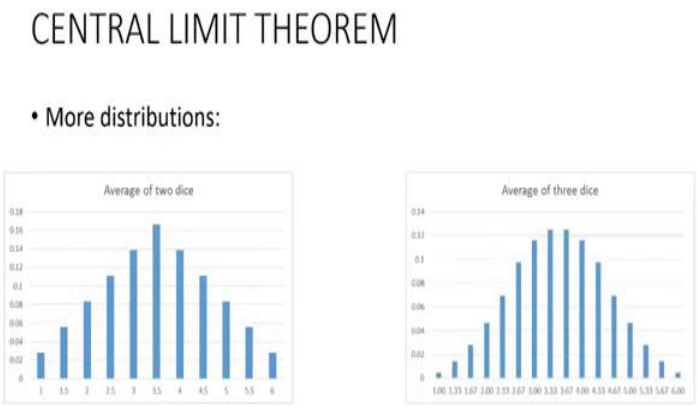
So, let us see how that works. So, on the left hand side of graph or here, I talk about the distribution associated with the single row and this view seen and we have discussed this is uniformly distributed. Why? Because, the heights are all in the same, which is discrete distribution and you see, it is uniformed distribution and it is 1 by 6; that is what I have shown here today. On the right hand side, I show you the distribution of the sum of two rows.

So, you can think of it is, rolling it once writing it down rolling it second time. So, you can think of your hands having you know two dice and you roll both of them and you sum up, what you see and what shows up. And already you can see that the distribution is started moving from uniform to something else. This happens to be triangular, but that is just the first step towards starting to look more and more bell shape ((Refer Time: 16:02)).

What is happening? Now, although the probabilities of rolling 1 through 6 were uniform, the summations; however, are not equal. So, the probability of getting a 2 is lower than the probability of getting a 3 and that should be fairly intuitive. For you to get a 2, you need to roll a 1 the first time and roll a 1 the second time. But, there are many ways in which you can get 3, mainly 2.

You can roll a 1 the first time and then, the roll a 2 or you can roll a 2 and then, roll a 1 and that kind of keeps increasing till you hit the point at 7, where 7 you can get in so many ways, you can roll a 6 the first time and then, roll a 1 the second or if you not, if you thinking of rolling both of the same time you can get a 6 and 1 or 1 and 6, 3 and 4, 4 and a 3 or 2 and a 5. So, there are more ways of achieving the same thing of achieving a 7, there are fewer ways of achieving a 2 or 3 and so, you already have something that is looking more like a normal.

(Refer Slide Time: 17:17)

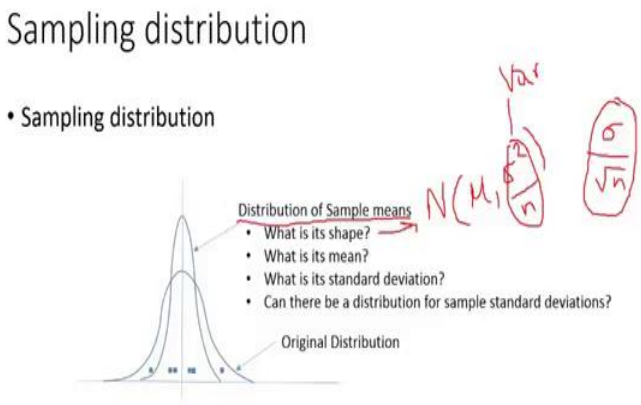


Now, you go further as I discussed; obviously, the average of two dice is the same as the sum of two dice. So, these two graphs ((Refer Time: 17:27)) are identical, this one and the next one on the next slide. These two are identical, except that are changed to average, so this axis is different. It goes through 1 through 6, the other one went from 2 to 12, but these are essentially identical graphs and this is also a triangular distribution.

But, look it, what is already happening. Now, if I say the average of three dice, so I am going to roll three dice at the same time or I am going to roll one after the other after the other. There are all independent either way. What you going to see is that, now this is started looking a little bit more you know triangular, let us starting to get that little bit of inflection and so on. And, so as you increase this number more and more, as the idea I said you get something that looks fairly normal and that is about the central limit theorem is about, that you aggregate a sufficiently large number of distributions when you start getting a normal distribution.

Now, this is the really important point for what we are going to say in next associated with sampling distributions. So, we going to start a fresh and sampling distributions, but I just want you to keep in mind, what we have discussed now in central limit theorem.

(Refer Slide Time: 18:44)



So, jumping give us, now to sampling distributions the idea here is very simple. So, lets the you have some original distribution and lets for now, say this distribution is normally distributed. And let us say this normal distribution has some mean, which I have shown with this blue vertical line and lets call that mu, so this point is mu. And let us say it has some standard deviation I am just referring to the dispersion through the arrow that is not the exact link of the standard deviation.

But, it is have some standard deviation, which can be represented is the variance is represented as sigma square and by the way this m mu comma sigma square is fairly norm nomenclature that just means it is a normal distribution with that looks like a with mu and sigma square all though that looks like an m, so may be a little bit more like mu norm. So, you have this distribution, now let us say N let us give it a name, so let us say this is the distribution of heights this is the distribution of what we lets keeps weights.

So, this is the distribution of weights staying consistence with the previous example of the men or the male members, who registered for introduction to data analytics. So, may be this distribution starts somewhere at I do not know 50 kgs and goes all the ways to say 100 kgs this is this is this is the distribution technically it can go all the way to infinity. Because, by definition and normal distribution can go to infinity and on this side it can go to minus infinity. So, this is this is the normal distribution.

Now, let us say that I took a sample from this distribution. So, these data points represent

the different samples and in this particular case I have taken just six samples, but well that can be more and the heights mean nothing the sample just mean, where they fall on the distribution. You can; obviously, use that and build a histogram and the idea is that if you build a if you take a sample large enough that histogram will fit very neatly to this curve, which is the normal distribution if that sample is very large.

If the sample is not you might get a different histogram, but what we most interested in is taking this sample and computing some key statistics from this sample. For instance if you took this sample and computed the arithmetic mean of the samples you will take each data point right and let us say you call it x 1 and the next data point got called x 2 and so on. Then, what you looking at is x 1 plus x 2 plus dot, dot, dot divided by N that is your arithmetic mean and, so you compute an arithmetic mean.

But ,since you got some finite sample got like 6 points and may be you have in an others instance 10 points the question is will your arithmetic mean always be equal to mu remember mu was, what defined this distributions this distribution is by definition mu comma normal mu comma sigma square. But, if you take a sample if you take some axis and compute x bar we differentiate between mu and x bar meaning mu is the theoretical mean, where as x bar is the sample mean it is.

If you take a sample of size n and compute an x bar will this x bar be equal to mu and both intuitively another wise the answer is no theoretically if your sample size is equal to infinity; that means, you take infinite number of samples, then perhaps your sample means will be equal to, then your sample mean again an theory will be equal to mu. But, that is not a practical situation, who takes infinite samples like that that by definition does not does not make is not very useful.

So, if you take a finite sample and in this case 6 and another case can be 20 next less say 20 and you compute a sample mean it is not going to be equal to mu. But, the idea is that it might the idea is that it is also a random variable, what do you mean by that you mean that say I mean one time you go about you take a sample you take a sample of 10. Let us say and you take the mean of that sample you will get a particular value that will not be equal to mu it could be equal to mu.

But, you know it could be little less than mu little greater than mu now, you go do that exact same thing again you will get some other new value. So, what; that means, is you

have a random variable on your hands and the random variable is about the distribution of the sample means for a given size end. So, that is what that is a core idea associated with sampling, which is that from the original distribution you take a sample and you compute a means and you get a certain value and, but that value itself belongs to a distribution that distribution changes based on the sample size.

So, suppose we were like I said if you took infinite if your sample size was really large if it was infinite, then perhaps you will not even have a distribution you just have a line out here which is that you almost always get mu because your sample size is, so large. But, if you , but think of the other extreme suppose your sample size is equal to one that is each time you took one point from the distribution and you computed the mean of that point, what is it means to compute the mean of a single points it is that number itself.

So, let us say we were looking at 50 kgs to a 100 kgs you took a random sample of one. So, 75 you know 65 kgs this time that was the random number I picked the average of 65 is 75. So, if you had a sample size of one what could the distribution of sample means look like the answer is it would look exactly like this distribution, because you taking a sample size of one its essentially like and you computing the average of that, which is nothing but, that number itself.

So, it is essentially like just re plotting that graph, now if your sample size was greater than 1, but less an infinity, what happens is if your sample size as the sample size gets larger and larger you are dealing with the distribution step. Because, each time you take a sample of, let us say 5 or 10 or 20 you are going to get some sample mean from that and that sample mean is not going to always be equal to the exact overall population mean.

And, but it is going to be some number nearby and the idea is that as in this particular case we had a normal distribution and the idea is that as long as we taking the average of a some number. Let us say 10 or 20 or 30 or 40 or 50 samples you are going to get a mean. But, that mean is not certain it is not certain, what that mean is going to be you know it is you know it need not be mu you know that for a fact.

So, what you are essentially getting is another distribution you getting a random number from another distribution and this the distribution of the sample means now it is. So, happens that when your original distribution is normally distributed the distribution of sample means is also normally distributed, but they might be some questions you have in

this regard. So, for instance what is the shape of this distribution the quick answer to the question is when the original distribution is normal like we said this distribution of sample means is also normal.

But, we also went through this central limit theorem where we said as long as your aggregating is sufficiently large number of distributions the resulting distribution starts to look normal. So, even if your original distribution is not normal as long as your aggregating a sufficiently large number this distribution of sample means becomes normal. So, that is the shape, now what is the mean of this distribution, what is the mean of the distribution of sample means the quick answer is because your just taking the average of some numbers if you what to do this is sufficiently large number of times you should not get a mean that is biased.

So, the mean of this distribution will also be equal to mu, but it is clear that the standard deviations are not the same right the standard deviation would be the same if your sample size was one in which, case you are not really sampling you are just taking a single data point. But, depending on the size of the sample the standard deviation is going to be typically lower a it will always be lower as long as the sample size is greater than 1 and the relationship is nothing but, sigma square.

So, if you are using sigma square it would be sigma square divided by N when you small enter refer to the sample size, but you can also think of it as taking the square root of this you can also think of it is sigma divided by square root of n. So, this would be the standard deviation and this would be the variance. So, this is v a r this is the variance and this is yours standard deviation.

So, that is the that is relationship that is very useful to remember, now you might have a question saying. So, we did all of this work to say let you have an original distribution you randomly sample from that distribution and you compute a mean an arithmetic mean then that arithmetic mean that you compute belongs has a distribution of its own and we spoke about the mean and shape and standard deviation.

Similarly, if you take a sample from the original distribution and you compute a standard deviation of that sample. Then, would you be is that sample standard deviation also coming from a distribution and the quick answer to that question is yes and in the if you are using a normal distribution to start with that is distribution of the sample standard

deviations tends to be chi square distributed and that is also something that we will encounter.

But, are focus for now has when on the distribution of sample means and the important things to take away are if you start with an original normal distribution, then by theory you will have a normal distribution for your sample means for whatever sample size. But, given that we also learnt about the central limit theorem even if you start with an original distribution that is not normal as long as you aggregate sufficiently large number of as long as your sample size is large enough and the distribution of sample means is likely to be normally distributed.

We spoke about how the mean of the distribution of sample means should be no different from the mean of the original distribution, because you are not adding or subtracting any number you just taking average numbers you are just taking numbers and taking the average of that. So, if you do that many times the distribution that you get from that should also be centered around the overall grand mean of the original distribution we spoke about how the standard deviation.

However, keeps reducing, so as long as you are aggregating more numbers your standard deviation will reduce in this rate it which, is reduces is as a function of this square root of n the sample size. So, sigma divided by square root of N is the rated, which the sample size your standard deviation of the distribution of sample means is with respect to the original distribution and actually that phenomena you should be able to see even in the examples that we took of the central limit theorem just two kind of show that you again see in this particular example and I will erase the red mark in this particular example I was focusing more and showing a central limit theorem about how the shape changes.

But, if you take this graph, which is this uniform distribution out here and there is some standard deviation out here right the sum spread around the mean correct this sum spread. Now, take a look at the average of two dice the mean is the same centered around the 3.5, but this spread has decreased right before this spread was like this. So, there was there was the higher probability of seeing values in the in the earlier graph up here you had data points that were with the higher probability further away from the center at 3.5.

Now, you do not see the probability of finding points far away from the center has

reduced the these are low probabilities, but the probability finding things close to the center is increased. So, therefore, the standard deviation of this distributions is lower than the standard deviation of the uniform distribution and that effect is going to just increase the probability of extremes keeps becoming lower there by the standard deviation becomes lower given that you are for all of these you are starting with one and ending with 6.

So, the that example shows both the central limit theorem meaning the change in the shape, but you can also capture this idea which is the distribution associated with sample means and in the previous cases the sample size was two in the first example and the sample size was three right because we were averaging two dice or three dice. So, the distribution that results from that is having a lower standard deviation.

So, that should give you an idea of the whole idea behind sampling distributions and this is the good concept to revise or understand deeply. Because, a lot of inferential statistics is based half of this and with that we conclude our lecture on random variables and probability distributions.

We will continue a next class and focus more on inferential statistics.