PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati

Week - 12

Lecture - 63

Notions of Convergence, Law of Large Numbers, and the Central Limit Theorem

Let us discuss the notion of convergence for a random variable and a sequence of random variables. Now, whenever we talk about convergence in a sequential sense, we know the sequence of real numbers. First of all, let us discuss the sequence of real numbers a_n. We say that this is a sequence of real numbers. So, it is basically a function from $\mathbb{N}$ (natural numbers) to $\mathbb{R}$ (real numbers).

For each value of n, we get a real number. So, it is known as a sequence of real numbers, denoted by a_n. So, we say that this sequence is a convergent sequence, and we denote it as: $\lim (n \to \infty)\ a_n = a$ if, for any $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $|a_n - a| < \varepsilon$ for all $n \geq n_0$.

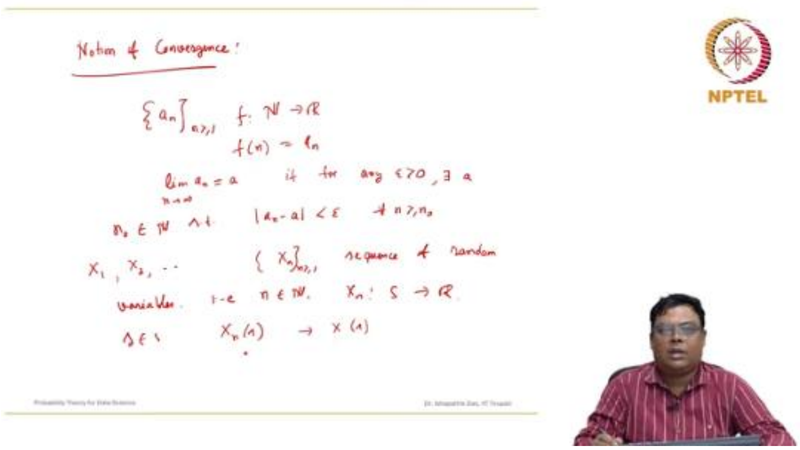
So, this is the concept of convergence in a sequence of real numbers. Now, if you consider a sequence of random variables, it is a sequence of functions.

For each n, we get random variables $X_1, X_2, \ldots, X_n$. This is a sequence of random variables. So, that means for each $n \in \mathbb{N}$, X_n will be a function from the sample space S to $\mathbb{R}$.

Now, how do we define convergence in this case? It's not a real number, and it's not just a function; it's a random variable.

This means it is a measurable function, but we can also talk about its probability. So, for some X_n, for any $s \in S$, $X_n(s)$ is a real number. Now, if you want to define that it is converging to some random variable X, how can we define this?

So, there are different approaches to define the convergence of a sequence of random variables. One is called convergence in probability.



So, a sequence of random variables X_n is said to converge in probability to a random variable X. X may also be a constant random variable, which we can say is some real number. In general, for convergence in probability to X, if for any $\varepsilon > 0$, we consider the difference $X_n(s) - X(s)$.

This difference $|X_n(s) - X(s)| \geq \varepsilon$ for all $s \in S$ such that $|X_n(s) - X(s)| \geq \varepsilon$. This will be a subset of S; it is an event. If you take the probability, then this will be a real number. Now, if this happens, n becomes a sequence of real numbers.

Now, if this goes to 0, it means we can write: $\lim (n \to \infty) P(|X_n - X| \geq \varepsilon) = 0$.

For simplicity, we will just write $\lim (n \to \infty) P(|X_n - X| \geq \varepsilon)$. If this probability goes to 0, then we say that for any $\varepsilon > 0$, $\lim (n \to \infty) P(|X_n - X| \geq \varepsilon) = 0$.

This is the same as the simplified notation. So, we can say that if this is equal to 0, then X_n converges to X in probability.

Now, let us discuss one example. In this example, suppose we consider $X_1, X_2, ..., X_n$ as independent, identically distributed (iid) random variables, where each has the same distribution, a normal distribution with mean μ and variance σ^2.

Now, if you change n, the number of random variables will change accordingly. Let us consider $\bar{X}_n$. We can express this as: $\bar{X}_n = (\Sigma X_i) / n$, which is the sample mean.

We know that the expected value of $\bar{X}_n$ is equal to μ, and the variance of $\bar{X}_n$, which we have already discussed earlier, can also be computed as σ^2 / n.

Let us compute it. The expected value of $\bar{X}_n$ is: $E(\bar{X}_n) = E((X_1 + X_2 + ... + X_n) / n) = (1 / n) \times (E(X_1) + E(X_2) + ... + E(X_n))$.

By definition, the expected values of $X_1, X_2, ..., X_n$ are all the same, which is μ, because all X's come from the same normal distribution.

So, $n \times \mu / n = \mu$.

Similarly, the variance of $\bar{X}_n$ can be found by taking the variance of $(X_1 + X_2 + ... + X_n) / n$. Since n is a constant, it will be: $(1 / n^2) \times (Var(X_1) + Var(X_2) + ... + Var(X_n))$.

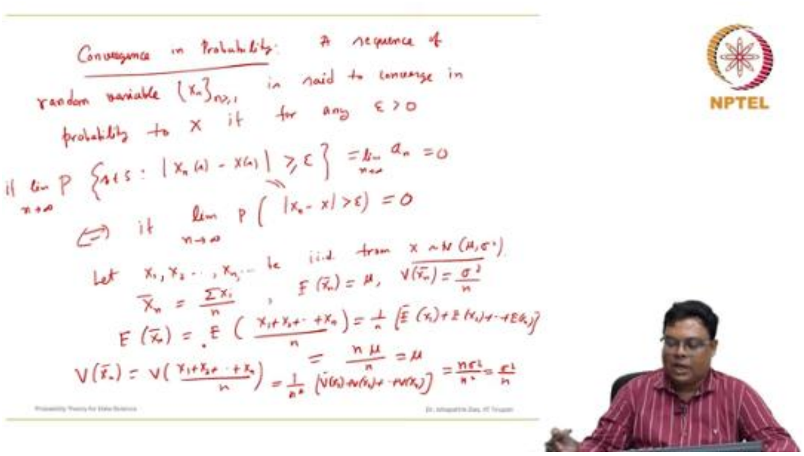
There will be some covariance terms, but since the variables are independent, the covariance will be 0.

So, all the variances will be the same, σ^2.

This becomes: $n \times \sigma^2 / n^2 = \sigma^2 / n$.

Now, the claim is that as $n \to \infty$, $\bar{X}_n$, which is a sequence of random variables for $n \geq 1$, converges to a constant random variable, μ, in probability.

How can we show that? Note that by Chebyshev's inequality, for any random variable, $P(|X - \mu| \geq \varepsilon) \leq \sigma^2 / \varepsilon^2$.



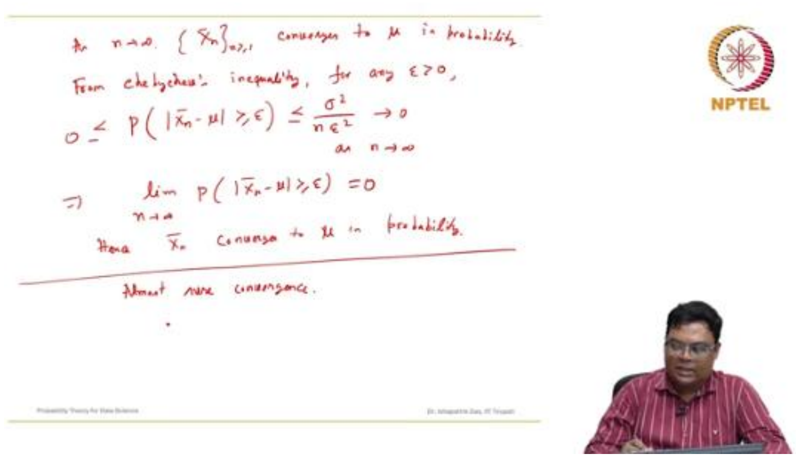
For any $\varepsilon > 0$, the probability that $\bar{X}_n$, which is the mean μ, is greater than or equal to ε is less than or equal to the variance of $\bar{X}_n$, which is σ^2 / n, and this is ε^2. Now, this probability is always ≥ 0, and it goes to 0 because it is just a sequence of real numbers as $n \to \infty$.

This implies that:
$\lim (n \to \infty) P(|\bar{X}_n - \mu| \geq \varepsilon) = 0$, by the sandwich theorem.

Hence, by definition, if this satisfies, we say that $\bar{X}_n$ converges in probability to X. So, $\bar{X}_n$ converges to μ in probability.

This is one notion of convergence. Another type of convergence is known as almost sure convergence. So, it is nothing but this function converges in pointwise limit as $n \to \infty$, with $X_n(s) \to$ [the limit].

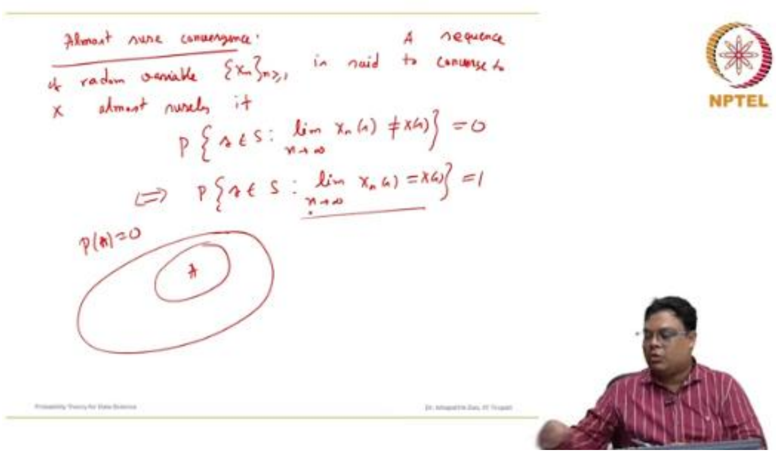
So, let us explicitly write that. This is almost sure convergence.



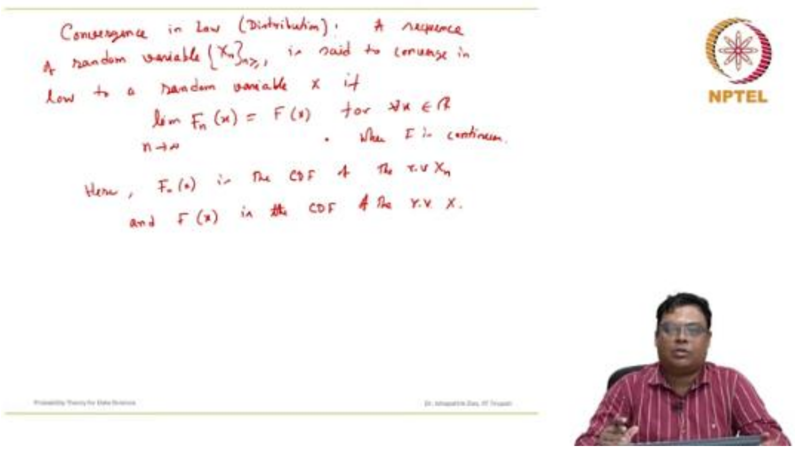
Let X_n be a sequence of random variables. A sequence of random variables, X_n, is said to converge to X almost surely if, for all s, the probability that the sequence converges approaches a limit as $n \to \infty$. Essentially, it converges pointwise for all $s \in S$ (the sample space). Now, it may not be converging for all s, but if the probability is equal to 0, it is equivalent to saying that the probability that for all $s \in S$, $\lim (n \to \infty) X_n(s) = X(s)$. If we observe that this probability is 1, then we say that X_n converges to a point X almost surely.

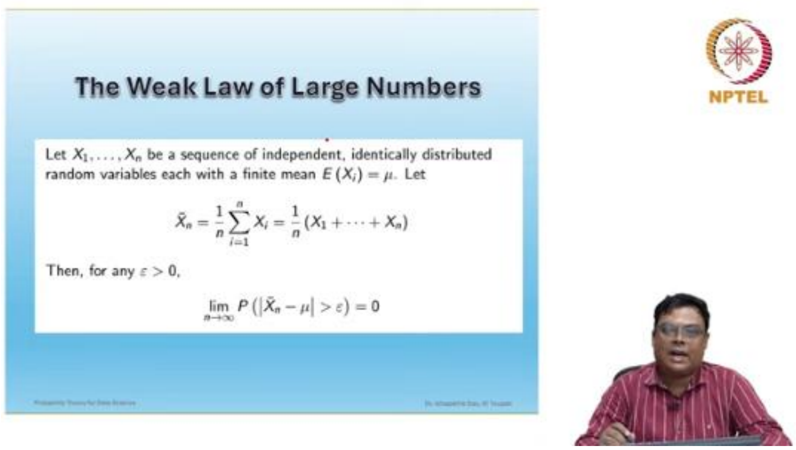
So, suppose this is a sample space, and you have some set A, where $P(A) = 0$. If the sequence of random variables converges at all points except for this set A, then it is called almost sure convergence. Almost sure convergence is like taking the majority, where there is some possibility or probability. If $P(A) = 0$, we will not consider those points. So, in a way, we're saying something similar to Lehman's idea, but the mathematical definition is that if $P(A) = 0$, and if the sequence of random variables converges at all other points where P is 1, then we say it is almost sure convergence.

Another notion of convergence is called convergence in law, also known as convergence in distribution. A sequence of random variables, X_n, is said to converge in law to a random variable X if, as $n \to \infty$, $F_n(x) \to F(x)$ for all $x \in \mathbb{R}$, where F is continuous. So, for all these X where F is continuous, if you observe, $F_n = F$. Here, the notation is that $F_n(x)$ represents the cumulative distribution function (CDF) of the random variable X_n, and $F(x)$ represents the cumulative distribution function (CDF) of the random variable X.



This defines the convergence in law. We will discuss some more examples in the following, whenever we learn some theorems or results and how they are utilized. First, let us discuss the weak law of large numbers. The weak law of large numbers states that we have already shown this kind of result. However, here we assume that convergence is in probability, with the mean (μ) and variance (σ^2) existing, and that the random variable X is normally distributed.



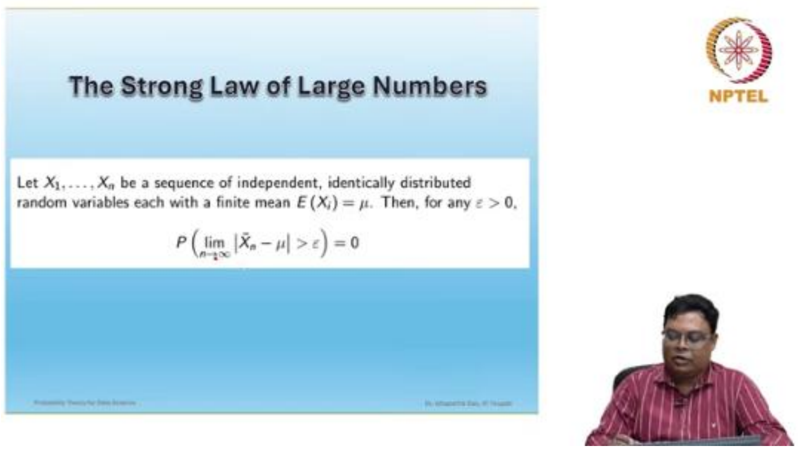


The weak law of large numbers is stronger than this concept. Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with a finite mean, μ. The variance is not specified, and $\bar{X}_n$ represents the mean of these variables. The weak law of large numbers states that as $n \to \infty$, $\bar{X}_n$ converges to μ in probability. This is known as the weak law of large numbers.

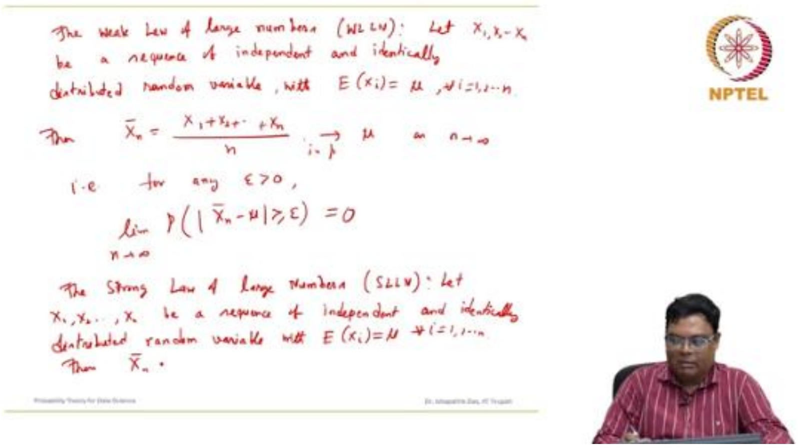
The weak law of large numbers states that... It is often abbreviated as the weak law of large numbers. Let $X_1, X_2, ..., X_n$ be a sequence of i.i.d. random variables. Since they are identically distributed, they all have the same mean, with $X_i = \mu$ for all i, where i ranges from 1 to n. Then it says that if you consider the sample mean of $X_1, X_2, ..., X_n$, this sample mean converges to μ in probability. This means that as the number of observations (n) $\to \infty$, for any $\varepsilon > 0$, the probability that $|\bar{X}_n - \mu| \geq \varepsilon$ approaches 0.

This is called the weak law of large numbers. We are not going to prove this here. The strong law of large numbers says the same thing, but we will not go into the details. So, this is just to learn it: the weak law of large numbers and the strong law of large numbers say that if $X_1, X_2, ..., X_n$ is a sequence of i.i.d. random variables with finite mean μ, then for any $\varepsilon > 0$, as $n \to \infty$, $P(|\bar{X}_n - \mu| \geq \varepsilon) \to 0$. Basically, this limit is inside here.

So, this applies to all s such that it is not converging to μ. Therefore, this is equal to 0. So, the strong law of large numbers says that the limit, whenever you apply it, remains the same. The statement for the strong law of large numbers is the same as the strong law itself. So, all these things will be the same.



Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables. Since they are independent and identically distributed, they all have the same mean, $X_i = \mu$ for all i from 1 to n. The Strong Law of Large Numbers states that $\bar{X}_n$, which is the sum of $X_1, X_2, ..., X_n$ divided by n, converges to μ almost surely as $n \to \infty$. This is almost sure convergence. In other words, you can say that $\lim (n \to \infty) P(\bar{X}_n = \mu) = 1$.



This is equal to 1. Alternatively, you can say that for any $\varepsilon > 0$, if you consider the probability lim ($n \to \infty$), the difference $|\bar{X}_n - \mu| > \varepsilon$ will be equal to 0. This means it is converging. So, this is the same as saying that $\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$. This is equivalent to saying these things.

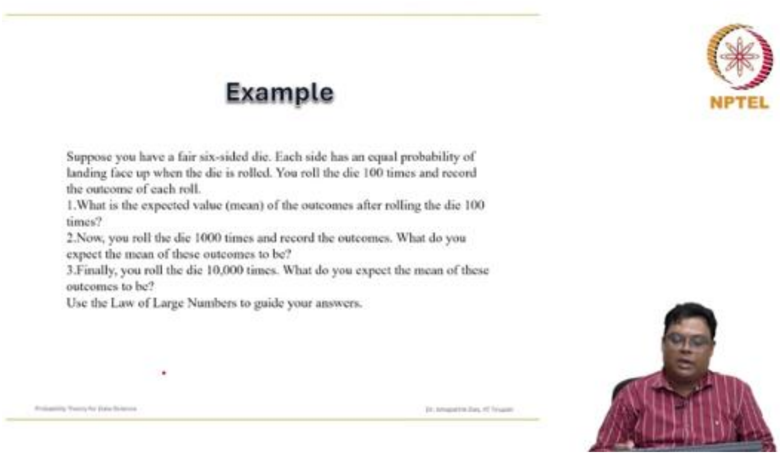
So, this is the Strong Law of Large Numbers. We won't go into more details about it, but it's important to mention both the Strong Law of Large Numbers and the Weak Law. So, why is it strong and why is it weak? Because almost sure convergence is stronger than convergence in probability. There are some examples where a sequence of random variables converges in probability, but it may not converge almost surely. But if a sequence of random variables converges almost surely, then it must converge in probability as well. That's why almost sure convergence is considered strong. So, that’s why it’s called the Strong Law of Large Numbers, and because it involves convergence in probability, it’s called the Weak Law of Large Numbers. We won’t go into more details here. Now, how is the Law of Large Numbers used?

There are many examples in real-life situations. For instance, if you go to a village during a Mela, you might see this in action. So, there are different types of gambling going on. It may seem like everything is happening randomly, and some people may be fortunate and make a profit. However, if you look at the theoretical part and how it's utilized, the Law of Large Numbers tells us that when n becomes very large, it's no longer random—it will approach a particular value.

So, the game is actually designed in such a way that the people running it—such as the shopkeeper or those who designed the game—will always make a profit in the long run. In the long run, they can compute the average behavior and the average profit. While individual trials may seem random and someone might win, in the long run, it is not random at all. It mostly converges to a particular point. For example, suppose you have a fair six-sided die, with each side having an equal probability of landing face up when the die is rolled.

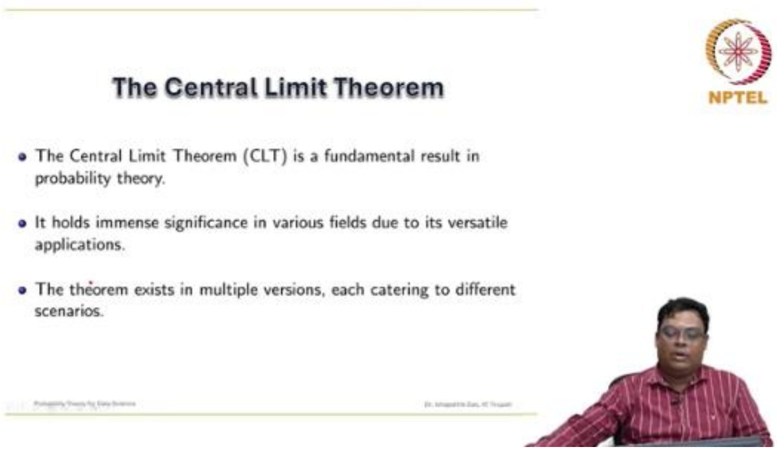
Now, if you roll the die 100 times and record the outcome of each roll, what is the expected value, or mean, of the outcomes after rolling the die 100 times? So, whenever you roll the die, you do not know what value it will land on. However, if you roll it 100, 1,000, or even 10,000 times, you can expect that each number will appear with a probability of 1/6. The number of times it should occur can be concluded based on this probability. We can discuss more details of this example later if we have time.

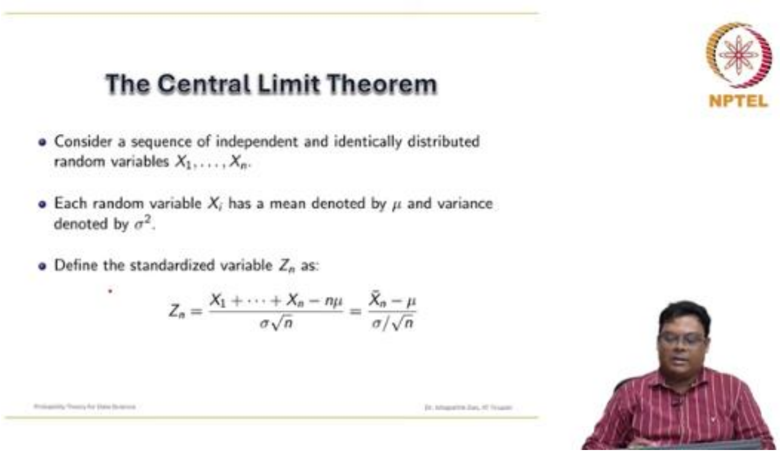
Then it will be very interesting to see how the Law of Large Numbers is useful in gambling, how it's used, and more details can be discussed with examples. Now, we will discuss another important topic. Often, when we collect data, we do not know the exact distribution of the data. So, it could be a Poisson distribution, binomial distribution, exponential distribution, or gamma distribution; we do not know. In that case, we assume it is a normal distribution.



The question then is, why are we assuming it to be a normal distribution? Why might it work or not? The basis for this assumption is the central limit theorem. The central limit theorem is a fundamental result in probability theory. It holds immense significance in various fields due to its versatile applications.

The theorem exists in multiple versions, each catering to different scenarios. So, here we discuss one version. Consider a sequence of independent and identically distributed random variables. There are many versions, but here we focus on these. Each random variable, X_i, has a mean denoted by μ and variance denoted by σ^2. Now, let us define Z_n.



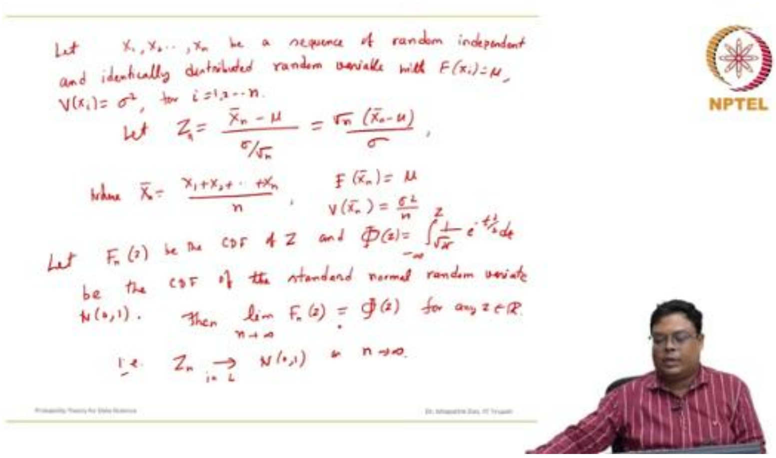


Let X_1, X_2, X_n be a sequence of random variables. So, consider a sequence of independent and identically distributed random variables. Here, we assume that each variable has a finite mean, as all have the same mean. The variance of each X_i is equal to σ^2 for all i from 1 to n, because the random variables are identically distributed, meaning their means and variances are the same. Let Z be defined as the difference between the average of the sequence of random variables, $\bar{X}_n$, and the mean, μ, divided by the standard deviation, σ, divided by the square root of n. If we simplify this, the square root of n is equal to the difference between $\bar{X}_n$ and μ, divided by σ. We have already defined $\bar{X}_n$ as the sum of X_1, X_2, ..., X_n, divided by n.

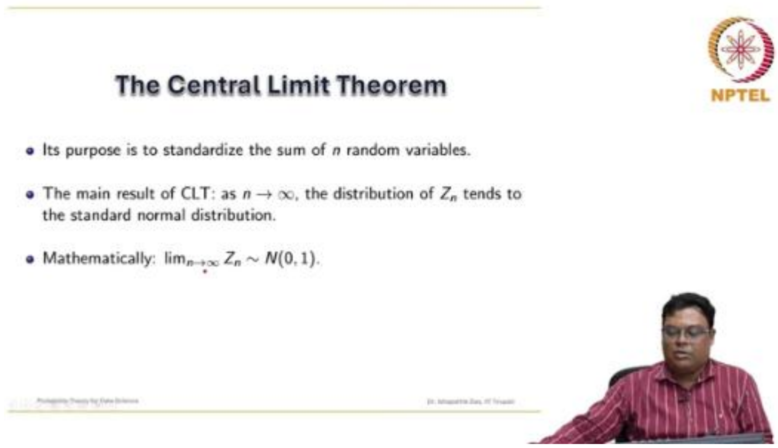
We have computed the mean of the expected values of $\bar{X}_n$, which is μ. The variance of $\bar{X}_n$ is σ^2/n. In this transformation, we are subtracting the mean and dividing by the standard deviation of $\bar{X}_n$, which is $\sigma/\sqrt{n}$. The central limit theorem states that, since it depends on n, let $F_n(z)$ be the cumulative distribution function of Z. We denote it as ϕ(z), which is defined by the integral from negative infinity to z of $(1/\sqrt{2\pi}) * e^{(-t^2/2)}$, with respect to t.

This is the cumulative distribution function of the standard normal distribution, or the cumulative distribution function of a normal distribution with mean 0 and variance 1, representing a standard normal random variate. So, we say that it is the standard normal random variate. Sometimes, we denote it as N(0,1), which represents a normal distribution with a mean of 0 and a variance of 1. This is referred to as the standard normal random variate. Then, as n approaches infinity, $F_n(z)$ for any z equals ϕ(z) for any $z \in \mathbb{R}$.

In other words, we say that Z_n converges to N(0,1) in distribution as n approaches infinity. The distribution function of $F_n(z)$ resembles the distribution function of the standard normal variate when n is very large. It is known as the central limit theorem. The central limit theorem has many applications. Here, the central limit theorem is given.



The purpose is to standardize the sum of n random variables. The main result of the CLT is that as $n \to \infty$, the distribution of Z_n tends to the standard normal distribution. Mathematically, as $n \to \infty$, Z_n is very close to N(0, 1). This is the central limit theorem. If you graphically represent this, suppose you take the transformation of Z_n, which is $(\bar{X}_n - \mu) / (\sigma / \sqrt{n})$.

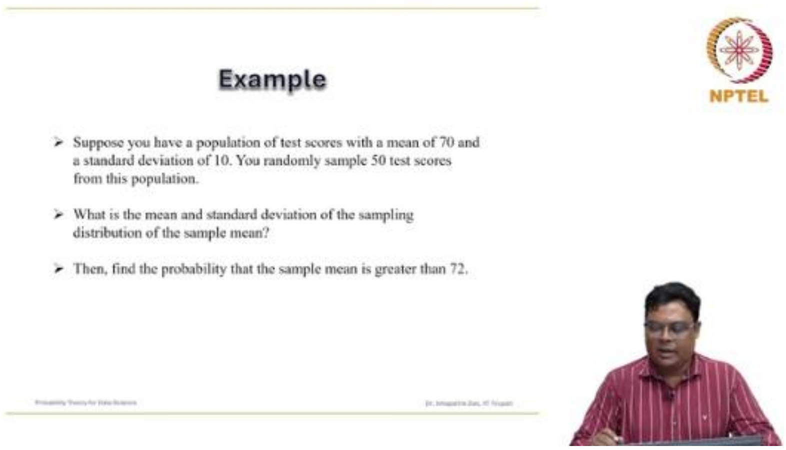


For small values of n, it may look something like this. So, when n is large, let's consider different values of n. Suppose n = 10, then n = 50, and n = 100. As n becomes larger, for example, when n = 1000, the distribution closely approximates a standard normal variate. This is the theorem.

So, when n is large, the distribution or probability density function of Z_n and its probability distribution function will closely resemble the standard normal distribution, which is denoted as N(0, 1). This is why, when n is large, we consider it to be a normally distributed random variable. This is the central limit theorem, and it has many applications. One of the applications is that the cumulative distribution function of $F\_Z_n(z)$ converges to the standard normal cumulative distribution function, $\varphi(z)$, as $n \to \infty$. This convergence is crucial for understanding the behavior of the sums of random variables.

Suppose you have a population of test scores with a mean of 50 and a standard deviation of 10. If you randomly sample 50 test scores from this population, what would be the mean and standard deviation of the sampling distribution of the sample mean? Additionally, find the probability that the sample mean is greater than 72. Let us discuss a numerical example where the central limit theorem can be applied. Suppose you have a population of test scores with a mean of 70 and a standard deviation of 10.

You randomly sampled 50 test scores from this population. Essentially, you have a sample of $X_1, X_2, ..., X_{50}$, meaning there are 50 test scores, with $n = 50$. These are independent and identically distributed (iid) random variables, all having the same mean because they are identically distributed, as stated by the population. So, from the population, you are drawing test scores with a mean of 70 and a standard deviation of 10. The variance, denoted as σ^2, is the square of the standard deviation, which is 100.



This holds for all i from 1 to 50. This information is given. Now, notice that the distribution of this population is not known. If the distribution is not known, whether it's a gamma distribution, an exponential distribution, or a normal distribution, we are unsure which distribution to use. The question here asks for the mean and standard deviation of the sampling distribution of the sample mean.

These can be determined because if you calculate the sample mean, denoted as $\bar{X}_n$, and here $n = 50$, it is simply the sum of X_1, X_2, and X_n divided by n. From here, we can compute that the expected value of $\bar{X}_n$ will be the same as the mean, which is 70. The variance of $\bar{X}_n$ will be σ^2/n. Since $\sigma^2 = 100$ and $n = 50$, the variance is 100/50, which equals 2. This is correct.

The variance of $\bar{X}_n$ can be found without using any specific distribution. We computed the expected value and variance of $\bar{X}_n$ earlier, and we did not rely on a normal distribution or any other distribution. You can see here that this calculation was done earlier. Whenever we discuss the law of large numbers, we can see that here. We can compute the mean and variance of $\bar{X}_n$.

Since the random variables are independent and identically distributed (i.i.d.), the variance of each X_i is σ^2. So, we computed the variance as σ^2/n. Now, we can calculate the value. As we have already found for this example, the expected value of $\bar{X}_n$ is 70, and the variance of $\bar{X}_n$ is 2. Now, what is the mean and standard deviation of the sampling distribution? The standard deviation of $\bar{X}_n$ will be the square root of the variance of $\bar{X}_n$, which is $\sqrt{2}$. So, the next question is to find the probability that the sample mean is greater than 72. We need to find this probability, where the sample mean $\bar{X}_n$ is greater than 72. Now, how can we find this? We do not know the distribution of $\bar{X}_n$.

Here, nothing is given about the population, only the mean and variance are provided. Now, if you use the central limit theorem, we know that if you consider Z_n, which is $(\bar{X}_n - \mu) / (\sigma / \sqrt{n})$, it approximately follows a normal distribution with mean 0 and variance 1 as $n \to \infty$, whenever n is large. Now, if we use the central limit theorem, we know that if we consider Z_n, which is $(\bar{X}_n - \mu) / (\sigma / \sqrt{n})$, it approximately follows a normal distribution with mean 0 and variance 1 as $n \to \infty$, whenever n is large. We can consider that since n will be large, it typically follows a normal distribution whenever n is greater than a certain value, such as 30. Most of the time, we can assume that it holds, but we must verify whether it is true.

Here, we assume that whenever n is large, it follows a normal distribution. Since $n = 50$ in this case, let us now consider $\bar{X}_n - \mu$. So, what is μ? $\bar{X}_n - \mu$ is 70, where $\mu = 70$. That is the value of μ, and $\sigma / \sqrt{n}$ is σ^2.

The variance of $\bar{X}_n$ here is the variance of the standard deviation. Therefore, the standard deviation of $\bar{X}_n$ is simply $\sqrt{2}$. We can replace this value here. So, this implies that the probability of $\bar{X}_n$ is the same as the probability of $(\bar{X}_n - 70) / \sqrt{2}$ being greater than a certain value. If you perform the same operation here, the inequality remains unchanged, and the probability will be the same.

So, this is essentially the probability that $(\bar{X}_n - 70) / \sqrt{2}$ is equal to a certain value. This might seem numerically complicated, but we will find out. So, this is Z_n, and it is greater than $(72 - 70) / \sqrt{2}$. So, this is equal to the probability that Z_n is greater than $\sqrt{2}$. Now, we know that Z_n is approximately normally distributed with a mean of 0 and a standard deviation of 1.

So, $Z_n > \sqrt{2}$ is approximately 1.414. This is the value for $\sqrt{2}$, and this is $-\sqrt{2}$. By symmetry, the probability that $Z_n > \sqrt{2}$ is the same as the probability that $Z_n < -\sqrt{2}$. Therefore, we will use approximate values of Z_n. The approximate value for $\sqrt{2}$ is $Z_n < -1.414$. I am using this value because we want to use this table. This table is the normal distribution table, and it already provides the negative values. So, the value for -1.41 is 0.07927, as given in the table to two decimal places. So, this is approximately 0.07927. You can see that the probability up to -1.41 is 0.07927. This is one application where, whenever n is large or when the distribution is unknown, we can use any distribution.

However, the best approach is to apply the central limit theorem. If n is large, it will mostly follow a normal distribution. By assuming that it is a normal distribution and using this transformation, we can find the required probability. I think we have completed the syllabus for probability theory for data science. I hope you have followed along and enjoyed this course.

You can refer to the books for more details and solve the assignments. By solving more problems, you can clear your doubts. I believe you have understood everything clearly, and I think this is the end of the course. Thank you.