PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati

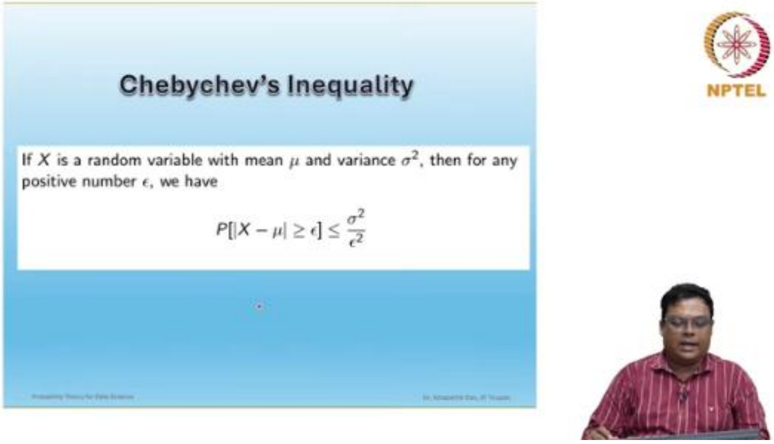
Week - 12

Lecture - 62

Chebyshev's Inequality

Let us discuss an important inequality known as Chebyshev's inequality. It has wide applications, and later we will explore some of these applications. It states that if X is a random variable with mean μ and variance σ², then for any ε > 0: P(|X - μ| ≥ ε) ≤ σ² / ε².

Let X be a random variable with mean μ and variance σ². For any ε > 0, P(|X - μ| ≥ ε) ≤ σ² / ε².



Chebychev's Inequality

If X is a random variable with mean μ and variance σ^2, then for any positive number ϵ, we have

$$P[|X-\mu| \geq \epsilon] \leq \frac{\sigma^2}{\epsilon^2}$$

Probability Theory for Data Science

Dr. Ishapathik Das, IIT Tirupati

In other words, this inequality provides a bound on P(X taking values outside the range [μ - ε, μ + ε]). This probability is:
P(|X - μ| ≥ ε) ≤ σ² / ε².

Now, let us prove this; it is not very difficult. Consider X as a continuous random variable, and then we will use integration. For the discrete case, the proof involves summation, but the approach is very similar.

Let X be a continuous random variable with the probability density function $FX(x)$.

Now, what is σ^2? By definition:
$\sigma^2 = \mathrm{Var}(X) = E[(X - \mu)^2]$,
where $\mu = E[X]$.

Therefore,
$\sigma^2 = \int$ from $-\infty$ to $+\infty$ $[(x - \mu)^2 * f\_X(x)]\ dx$.

This is the formula for variance. Note that this is the sum of positive numbers. Since $f\_X(x) > 0$ and $(x - \mu)^2 \geq 0$, all values in the integral are positive.

If we perform the integration over a subset of real numbers where $|X - \mu| \geq \varepsilon$, the integral will also be $\geq$ a certain amount, as we are adding positive terms without cancellation.

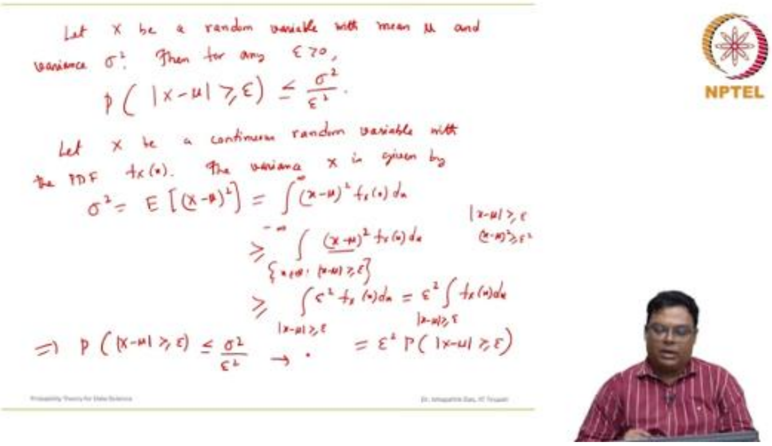
In this region, $|X - \mu| \geq \varepsilon$ implies $(X - \mu)^2 \geq \varepsilon^2$. Therefore,
$\int$ (over $|X - \mu| \geq \varepsilon$) $[(x - \mu)^2 * f\_X(x)]\ dx \geq \varepsilon^2 * \int$ (over $|X - \mu| \geq \varepsilon$) $[f\_X(x)]\ dx$.

Simplifying further:
$\sigma^2 \geq \varepsilon^2 * P(|X - \mu| \geq \varepsilon)$.

Rearranging:
$P(|X - \mu| \geq \varepsilon) \leq \sigma^2 / \varepsilon^2$.

This proves Chebyshev's inequality. The proof is straightforward; you can review it again if needed.

Now, we will see how this inequality can be utilized.



Let X be a random variable with mean μ and variance σ^2. Then for any $\epsilon > 0$,

$$P(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Let X be a continuous random variable with the PDF $f_X(\cdot)$. The variance X is given by

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx$$

$$\geq \int_{\{x \in \mathbb{R} : |x-\mu| \geq \epsilon\}} \underline{(x-\mu)^2} f_X(x)\,dx$$

$|x-\mu| \geq \epsilon$

$(x-\mu)^2 \geq \epsilon^2$

$$\geq \int_{|x-\mu| \geq \epsilon} \epsilon^2 f_X(x)\,dx = \epsilon^2 \int_{|x-\mu| \geq \epsilon} f_X(x)\,dx$$

$$= \epsilon^2 P(|X-\mu| \geq \epsilon)$$

$$\Rightarrow P(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Probability Theory for Data Science

Dr. Ishapathik Das, IIT Tirupati

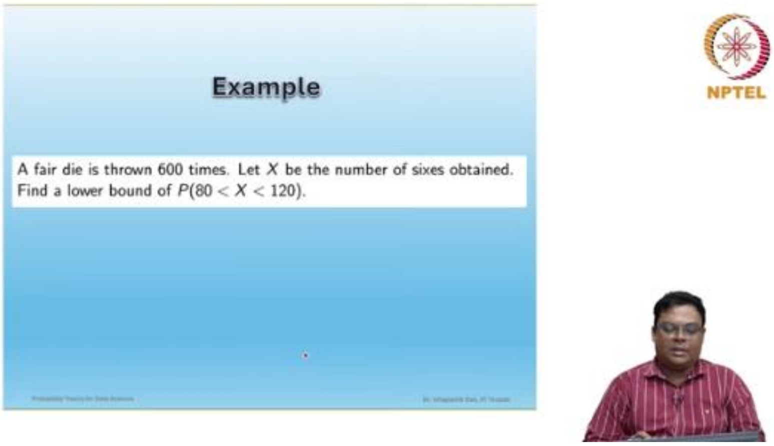
This inequality is important. Let us look at a numerical example:

A pair of dice is thrown 600 times. Let X = number of 6s obtained.

So, when you are throwing a pair of dice 600 times, X represents the count of outcomes where a 6 appears.

For a pair of dice, the probability of getting a 6 on a single die throw is: P(6) = 1/6.

The number of times a 6 can occur is related to P(X = 6). Since the dice are fair, P(6) = 1/6.



Example

A fair die is thrown 600 times. Let X be the number of sixes obtained. Find a lower bound of $P(80 < X < 120)$.

Probability Theory for Data Science

Dr. Ishapathik Das, IIT Tirupati

Now, suppose we consider another random variable, Y. This is the result after each die throw. X represents the random variable for the number of 6s. Specifically, for each individual throw, the probability of getting a 6 is 1/6. However, X represents the total number of 6s after throwing the die 600 times.

So, we have Y_i, where i ranges from 1 to 600. We know that this follows a binomial distribution with parameters $n = 600$ and $p = 1/6$, where 600 represents the number of trials and 1/6 is the probability of getting a 6 on each die throw. Now, we are asked to find the lower bound of this value. Let X represent the number of 6s. Therefore, $X \sim \text{Binomial}(n = 600, p = 1/6)$.

We need to find the lower bound for $P(80 < X < 120)$. So, we are asked to find the lower bound. Let's write down the problem: find the lower bound of $P(80 < X < 120)$. So, now X can be, sorry, $80 < X < 120$. This probability is always ≥ 0, but we are looking for a non-trivial lower bound.

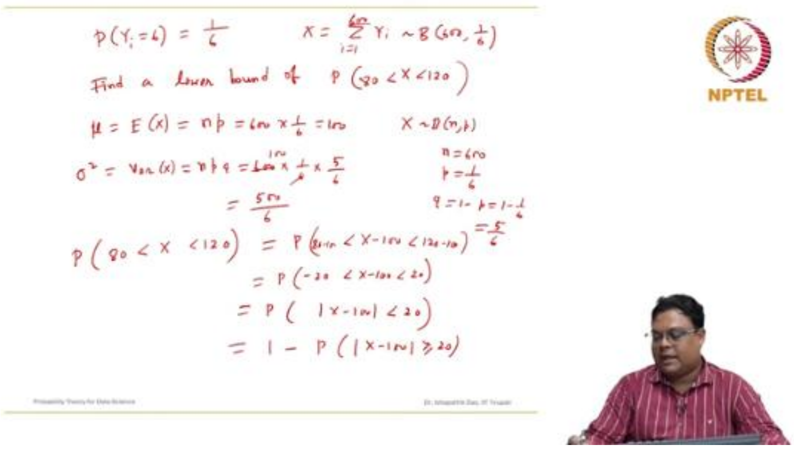
How can we find this lower bound? Now, let us find the expected value of X. So, the mean, denoted as μ, is equal to $E[X]$. We know that for a binomial distribution, $E[X] = n \times p$. In this case, $n = 600$ and $p = 1/6$.

So, $600 \times (1/6) = 100$. Now, what is the variance, σ^2? The variance of X is given by $n \times p \times q$. In this case, $600 \times (1/6)$, and $q = 1 - p = 5/6$. So, $q = 1 - (1/6) = 5/6$. Now, we can simplify this.

So, $600 \times (1/6) = 100$. Let's keep it as 500/6 for now, and we will simplify it later if needed. So, this is the value of σ^2. Now, Chebyshev's inequality says that $|X - \mu|$. So, what is $|X - \mu|$?

Here, we want $P(80 < X < 120)$. This is equal to $P(-20 < X - 100 < 20)$. Essentially, the probability is that $|X - 100| < 20$.

Now, this is equivalent to $1 - P(|X - 100| \geq 20)$. So, for this case, we have some device. By Chebyshev's inequality, we know that for any $\varepsilon > 0$, $P(|X - \mu| \geq \varepsilon) \leq \sigma^2 / \varepsilon^2$. Now, if you subtract both sides, we will get $1 - P(|X - \mu| \geq \varepsilon) \geq 1 - \sigma^2 / \varepsilon^2$.



$P(Y_i = 6) = \frac{1}{6}$ $X = \sum_{i=1}^{600} Y_i \sim B(600, \frac{1}{6})$

Find a lower bound of $P(80 < X < 120)$

$\mu = E(X) = np = 600 \times \frac{1}{6} = 100$ $X \sim B(n, p)$

$\sigma^2 = Var(X) = npq = \overset{100}{\cancel{600}} \times \frac{1}{\cancel{6}} \times \frac{5}{6}$ $n = 600$

$= \frac{500}{6}$ $p = \frac{1}{6}$

$q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}$

$P(80 < X < 120) = P(80 - 100 < X - 100 < 120 - 100)$

$= P(-20 < X - 100 < 20)$

$= P(|X - 100| < 20)$

$= 1 - P(|X - 100| \geq 20)$

Probability Theory for Data Science Dr. Ishapathik Das, IIT Tirupati

First, we can subtract this, and it will be ≥ 1 when we add 1. So, this is nothing but another version of the same thing. It is essentially the complement of $|x - \mu| < \varepsilon$, which is $\geq 1 - \sigma^2 / \varepsilon^2$.

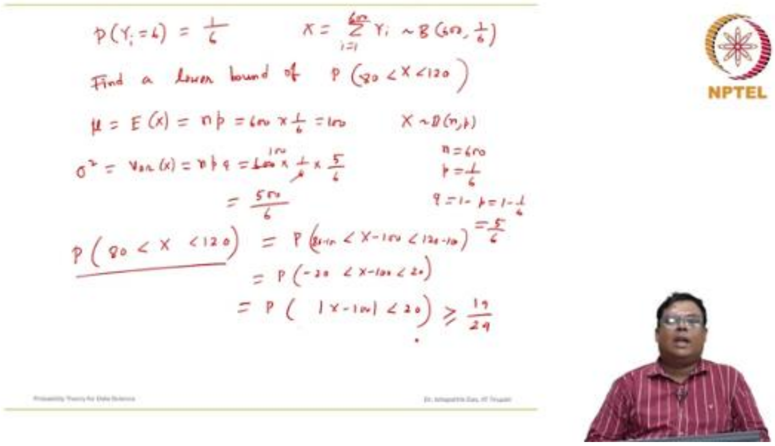
Now, here you can see that we want to find $\mu = 100$, and we want to know what this is. So, let's keep it as it is, because we can use these things from Chebyshev's inequality. This is another way of writing it, and we've also found the lower bound.

It's the upper bound for this case, and we did that for the complement part as well. Hence, we want $|x - 100| < 20$. So, $\varepsilon = 20$, and we also know that $\sigma^2 = 500/6$. Therefore, we can calculate the probability that $|x - 100| < \varepsilon = 20$. This is $\geq 1 - (\sigma^2 = 500/6) / (\varepsilon^2 = 20^2)$.

So, then we cancel this out. We get $1 - 5/24$, which simplifies to $19/24$. Hence, what do we have finally? We were looking for the lower bound of this probability, and that is the answer. This is the lower bound of the probability.

This probability is the same as this probability, and the same as this probability. Finally, what did we find? Using Chebyshev's inequality, we found that it is $\geq 19/24$. This is one application of Chebyshev's inequality. We will discuss more applications later, as we continue learning.

Most of the parts of this course have been completed. The remaining topics we will discuss include the notion of convergence. Some of the random variables and related laws, such as the law of large numbers, will be discussed. We will also cover the strong and weak laws of large numbers, along with the central limit theorem. The course will be completed with additional numerical examples.



$P(Y_i = 6) = \frac{1}{6}$ $\qquad X = \sum_{i=1}^{600} Y_i \sim B(600, \frac{1}{6})$

Find a lower bound of $P(80 < X < 120)$

$\mu = E(X) = np = 600 \times \frac{1}{6} = 100$ $\qquad X \sim B(n, p)$

$\sigma^2 = Var(X) = npq = \overset{100}{\cancel{600}} \times \frac{1}{\cancel{6}} \times \frac{5}{6}$ $\qquad n = 600$

$p = \frac{1}{6}$

$= \frac{500}{6}$ $\qquad q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}$

$\underline{P(80 < X < 120)} = P(80-100 < X-100 < 120-100)$

$= P(-20 < X-100 < 20)$

$= P(|X-100| < 20) \geq \frac{19}{29}$

NPTEL