

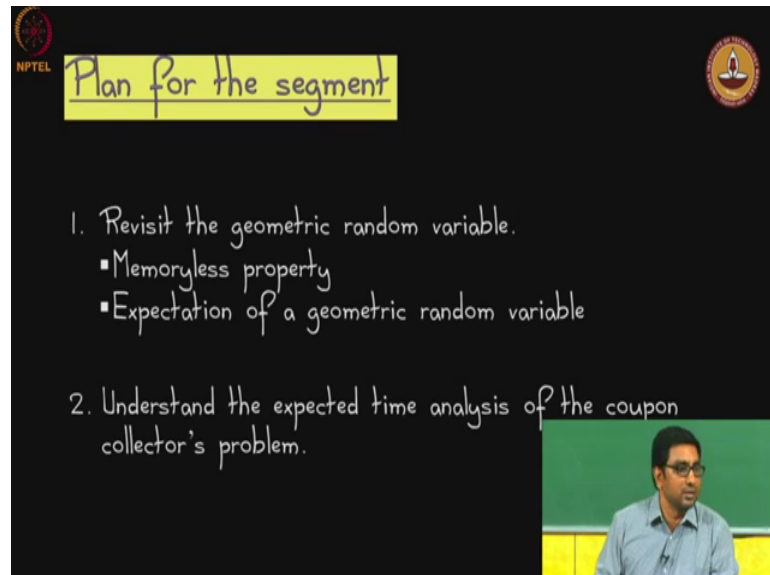
Probability & Computing
Prof. John Augustine
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 02
Discrete Random Variables
Lecture - 13
Discrete Random Variables – Geometric Random Variables & Collecting Coupons

Let us get started. We are in 5th segment of module 2 where we will talking about random variables, and in particular view we talked about the expectation of random variable and the conditional expectation.

Today I mean this segment on the next segment we are going to talk about some algorithmic ideas, so let us hopefully going to change um. So, we are going to talking about geometric random variables and then will apply them to understanding this problem call coupon collectors problems that is that is going to be the topic for today.

(Refer Slide Time: 00:47)



The slide is titled "Plan for the segment" and contains the following content:

- 1. Revisit the geometric random variable.
 - Memoryless property
 - Expectation of a geometric random variable
- 2. Understand the expected time analysis of the coupon collector's problem.

A small video inset in the bottom right corner shows Prof. John Augustine speaking.

And of course, we have already seen with geometric random variables. So, that sense we will just revisited and we will understand one important property of this random variable called the memoryless property and then we will work out the expectation understand the coupon collectors problem and analyze its expected time ok.

(Refer Slide Time: 01:15)

The Geometric Random Variable with parameter p - Revisited

- Number of flips of a coin with bias $p < 1$ until heads
- Geometric random variable X has support $1, 2, \dots$
For $i = 1, 2, \dots$

$$\Pr(X = i) = (1 - p)^{i-1} p$$

$\sum_{i=1}^{\infty} \Pr(X = i) = 1?$

$$\begin{aligned} \sum_{i=1}^{\infty} \Pr(X = i) &= \sum_{i=1}^{\infty} p(1 - p)^{i-1} \\ &= \frac{p(1 - (1 - p)^{\infty})}{1 - (1 - p)} = 1 \end{aligned}$$

So, let us revisit the definition of the geometric random variable. So, this the best way to illustrate this is it is the number of flips of a coin with some bias p until you get the first heads ok. So, this geometric random variable comes with the parameter p and that shows up in this definition this intuitive definition.

For formally if you have a geometric random variable X it has support of $1, 2, 3$ and so on all the integer starting from one when I say support what does; that means, is these are the values for which the probability is non-zero. So, for i equal to value ranging from one onwards what is the probability that the random variable X takes the value i ; that means, the previous i minus 1 flips must been heads and that is why you have $1 - p$ this is the probability that you get tails raise to the power i minus 1 followed by the probability that you get a heads.

So, this is the formal definition of the geometric random variable. So, if it is proper distribution you need to have this property that the probability of the sample space should equal 1 ok, and thus if we can verify that and so that you can verify just by summing over all possible elements in the sample space probability that X takes that value corresponding to that i .

And so now, we apply the definition which we already seen before then notice that you can this is bah essentially let us see what this is over here its basically the sum of a

geometric series if you apply that formula you end up getting know. So, that fits our requirement that the probability of the sample space must be a 1.

(Refer Slide Time: 03:18)

Memoryless Property

$$\Pr(X = i + k \mid X > k) = \Pr(X = i)$$

Proof.

$$\begin{aligned} \Pr(X = i + k \mid X > k) &= \frac{\Pr((X = i + k) \cap (X > k))}{\Pr(X > k)} \\ &= \frac{(1-p)^{i+k-1}p}{\sum_{j=k}^{\infty} (1-p)^j p} = \frac{(1-p)^{i+k-1}p}{(1-p)^k} \\ &= (1-p)^{i-1}p = \Pr(X = i). \end{aligned}$$

Then comes the very important property memoryless property. This is very crucial property it is a sometimes can be little unintuitive and what this means is that let us say the let us let us go back to the tossing of coins way of looking at this distribution. If you toss the coin of few times and you been repeatedly getting tails, then let us say you have done this for 5 times in repeatedly got the 5 tails.

And let say then that is history have any bearing on how many more coin tosses you will need before you get the heads and as you turns out you history will not have any bearing care because these are independent coin flips and that is the intuition of this memoryless property is capturing. The previous coin tosses if you are unlucky enough you gotten tails will not somehow influence you to get a heads quickly ok.

And this is some something that goes counter to a lot of our thinking because you know, people talk about things like you know oh I had the you know I had the good things happen to me and now worried about something bad happening or if bad things happening to me I thought to be getting a good thing you know sometime soon, yeah. You know if this out of memorylessness property shows up in life as well then that will not be the case. I do not know whether it shows up in life or not, but in this distribution it is completely memoryless case ok.

So, let us formally see why that is the case. So, how do we express that formally. So, we are asking what is the probability of X equal to some i plus k given that X is greater than k what does that mean. So, for the first k coin tosses your head tails you will be observed that given that you know that X therefore, has to be some value greater than k ok. So, what is this conditional probability on the left hand side? Now, the right hand side look there has no k is completely eliminated. So, the fact that you seen k is coin tosses is completely eliminated on the right hand side this is the probability that X equal to i .

So, let us see why this is true in a formal way umm. So, the left hand side we take it and we just apply the formula for conditional probability and then so on the in the numerator is what do we have is the probability that X equal to i plus k intersected with the event that X is greater than k ok. So, if X is equal to i plus k and i and k are both positive quantities it is clearly what you have on the numerator is basically just what happens you have to get i plus k minus 1 tails followed by the heads and X has to be greater than k ok. And in the denominator we have a prox; so that is the numerator in the denominator let us see little bit careful. So, what you are saying is the probability that X is greater than k .

So, if X is greater than k it can take the value k plus 1, k plus 2 and so on and you have to sum it over all those possibilities and so we run it through a summation starting from j equal to k to infinity. So, over the first j coin tosses have to be tails followed by a heads. Of course, you have to p 's in the one in the numerator and one in the denominator. So, they will cancel out.

And one thing I would like you to work it out on your own is basically the this summation if you work through it is going to end up being $1 - p$ raise to the k divided by p . So, it will come out this way in this derivation. So, when you work it out it is going to be. So, what happens over here $1 - p$ raise to the k will cancel out with this k over here. So, you will get $1 - p$ raise to the i minus one times this p which is nothing, but the probability that X equal to i . So, this is formally verifying our intuition of the memoryless property.

(Refer Slide Time: 07:43)

$E[X] = \frac{1}{p}$

• Let $Y = 0$ when first flip is tails. $Y = 1$ otherwise.

$$E[X] = \Pr(Y = 0)E[X | Y = 0] + \Pr(Y = 1)E[X | Y = 1]$$
$$= (1-p)E[X | X > 1] + p$$
$$= (1-p)E[X + 1] + p$$
$$= E[X] - pE[X] + 1 - p + p$$

Thus,

$$E[X] = \frac{1}{p}$$

So, now let us look at the geometric random variable and let us ask what is the expectation of the geometric random variable. We want to be we want to claim that its 1 over p and this should make into two decision. So, now, let us say that this geometric random variable the coin has bias very small bias; that means, it is going to take more coin tosses to get the first head.

So, say bias is one-tenth ok. So, roughly only at tenth of the coin tosses you are going to see is heads. So, you will have to toss roughly ten times before you see the first head and that is intuitive statement and so when we try to formalize that we will be able to we will state that has expectation of X equal to 1 over p . So, let us see why that is exactly correct. So, when you think of let us focus on the first coin flip. The first coin flip can either be a tails or a heads. So, let us say that define another random variable Y it is as bernoulli random variable. So, just X is equal to 0 if the first flip is tails and one is the first flip is a heads and so now, the expectation of X you can write it in this form in this fashion.

So, now let us say there are two possibilities, either the first flip is the tails or the first flip is a heads, and this might be will a little bit easier. So, when the first flip is a is a basically let us actually look at this line, depending on whether you get tails or heads the expectation becomes conditional on that. So, the expectation of X conditioned on Y equal to 0 here, Y equal to 1 here. Let us actually see how that plays out.

Now, we will be easier to see this part. When Y equal to 1 what is that mean? It just means of the very first coin flip was a heads and this probability itself is p and this then this conditional expectation becomes a 1 because in very first flip you got a heads ok. In this part the probability is $1 - p$, but what about the conditional expectation. Well, when you say Y equal to 0 it means the first flip failed you cannot have a value of X equal to 1 for this in this case. So, we can condition instead of conditioning on Y we can condition on X mean greater than 1, because Y X equal to 1 is out of the question now.

So, now what is this expectation of X given X greater than 1? When you think about it now we are applying the memoryless property, X when its greater than one its what when it is when its guarantee there is greater than 1 has to be at least one the first flip has to be a tails, after which entire memory is lost and you are basically it is like starting the experiment all over again ok. So, this one is counting for the fact of first coin flip was a tails and it is completely lost you have to restart the experiment. So, then you have to add this X , so this X given X greater than 1 can be written as just express 1.

Now, we can apply linearity of expectation. So, and I am skipping a step here linearity of expectation and then multiply with $1 - p$ you will get p 's terms and you will get a few cancellation. So, p and minus p will cancel out and so here what let us actually what this of (Refer Time: 11:32) bit carefully. So, this E of X will cancel out the this e of X this p will cancel out the this p and so what you will have is p times E of X equal to 1 which is which will then gives give this form.

So, this again confirms our intuition that the expectation of geometric random variables 1 over.

(Refer Slide Time: 12:01)

The Coupon Collector

Given: A collection of n coupons

repeat
Sample a random coupon
until all coupons sampled

How many iterations?

The slide features a dark background with yellow and white text. On the right side, there is a diagram of a horizontal rectangle divided into three equal-sized empty boxes. Five orange arrows point towards these boxes: two from the top and three from the bottom. In the bottom right corner, there is a small inset video frame showing a man with glasses and a light blue shirt speaking in front of a green chalkboard.

Now, comes this very interesting problem the coupon collectors problem the way to it is sort of explain this is at least as a fun way to think about it. So, let us say you are buying some something in the store and each time you buy you get a nice sticker ok.

And let us say there are some n different types of stickers, and you want to in each time you want to buy this box of chocolates or whatever you will get this sticker and you get a random sticker out of the n different stickers of the company has made available. And so you are asking how many time should I buy this box of chocolates before I get at least one copy of all the n stickers.

So, let us state that formally, you are given a collection of n coupons if you were stickers coupons whatever you want to call it. And then what you do? You when you buy this box of chocolates you get one of them ok. So, think of it is sampling a random coupon and you want to now repeat this process until you gotten all the possible coupons at least once ok.

So, maybe the next time you buy you get this coupon, and the next time you buy unfortunately you get something that you have already seen ok. And next time you get something new ok, and then you again get unlucky you get something you are already seen before and finally, you get to see something the last coupon. So, this point you seen all the 4 different coupons right. So, that is the that is the coupon collectors problem. And the question is how many times should we buy the box of chocolates before we get to see

all the stickers or another way to stating it is how many iterations of this procedure here should be executed before we have gotten all the coupons. And this is a simple problem that shows up in a lot of sampling situation. So, it is important to understand this, this has been this can be analyzed quite thoroughly, but for now we are going to just focus on understanding the expected number of iterations ok.

(Refer Slide Time: 14:21)

The slide features a black background with white and blue text. At the top left is the NPTEL logo, and at the top right is the IIT Bombay logo. The title 'The Coupon Collector's Problem' is written in blue. Below it, the theorem is stated: 'Theorem. Let X be the number of iterations of the coupon collector's problem. Then, $E[X] = n \ln n + \theta(1)$.' The word 'Proof.' follows, and the definition of X_i is given: 'Let X_i be the number of iterations after $i - 1$ different coupons seen until a new (i th) coupon is seen'. A yellow thought bubble contains the text 'Oops, replace 1 with n.' In the bottom right corner, there is a small video inset showing a man in a blue shirt speaking.

So, this is the theorem we want to prove. That X be the number of iterations of the coupon collectors problem. So, the number of times you will sample the expectation of X is equal to some n times \ln of n plus a smaller n term. So, how does this proof go we do this sort of breaking up of X . So, we break up this X into small x size ok. So, X in particular X_i is the number of iterations after you seen i minus 1 different coupons, but until you see the i th coupon ok.

So, let us make sure we understand what this means ok.

(Refer Slide Time: 15:09)

The Coupon Collector's Problem

Theorem. Let X be the number of iterations of the coupon collector's problem. Then,
$$E[X] = n \ln n + \theta(1).$$

Proof.

Let X_i be the number of iterations after $i - 1$ different coupons seen until a new (i th) coupon is seen

$X_1 = 1, E[X_2] = \frac{n}{n-1}, \dots, E[X_n] = n.$

The slide also features a timeline diagram with vertical bars and a small video inset of a man speaking.

So, this is let us say the timeline you are sampling overtime the very first time you buy something you are going to see something new. So, your X_1 basically is have to seen 0 different coupons until you see the first new coupon. So, that is this X_1 equal to 1 ok. Very first time we buy something we will get something new. The second time use, so at this point in time you using one coupon there are n minus 1 coupon that you have not see ok.

So, now you ask what is how many time should I buy before I see one more new coupon that is going to be your X_2 ok. So, what is the expectation of X_2 ? You think about it its now going to be a geometric random variable, your success, your p what is basically n minus 1 over n because there are n minus 1 coupons you have not seen before out of a total of n and if you get any one of them you have seen a new coupon that is your p value. And what is the expectation of X_2 ? That is 1 over p . So, that is n over n minus 1.

And similarly X expectation of X_3 if you work it out its going to be n over n minus 2 and so on and so the pattern will continue on ok. And this should fit your intuition because an early on it is the these quantities are going to be very close to one expectation of X_2 is going to be closed to one expectation of X_3 is going to be close to 1 and so on ok. And this should fit your intuition because early on its going to easy to find new coupons ok, but as you start collecting coupons is going to get harder and harder to see

new coupon because every time you buy you are going to find the coupon it is likely that you are going to find the coupon that you already collected.

And particular if you look at the very last X_n is going to an expectation take n time before you find that coupon because you seen $n - 1$ you only have one coupon that you have not seen out of a total of n . So, your p reduces to 1 over n . So, the expectation becomes 1 over p which is equal to n ok, and that should fit your intuition.

(Refer Slide Time: 17:48)

The Coupon Collector's Problem

Theorem. Let X be the number of iterations of the coupon collector's problem. Then,

$$E[X] = n \ln n + \theta(n).$$

Proof contd.
 Clearly, $X = \sum_{i=1}^n X_i$, so

$$E[X] = \sum_i E[X_i] = \sum_i \frac{n}{n-i+1} = n \sum_i \frac{1}{i}$$

$$= n \mathcal{H}_n = n(\ln n + \theta(1))$$

$$= n \ln n + \theta(n).$$

So, now, let us know now that we know the expectations let us plug them into our understanding of X . So, clearly the capital X is this is this is just breaking few clearly by just breaking time into X_1, X_2 and so on up to X_n right. So, capital X simply the summation of these X_i 's and we can apply the linearity of expectation and apply the formula for E of X_i and. So, that is going to be summation over i n over $n - i + 1$, and n is common you get it out and summation i^2 over i .

What is summation i^1 over i ? That is nothing, but the n th harmonic number and we have a formula for that it is the textbook goes through the details of how these formulas arrived at. But we will skip those details, but essentially n th harmonic number is nothing, but it is between $\ln n$ and $\ln n + 1$ and so you can write that as $\ln n$ plus $\theta(1)$ and with that we get the result; that means, 1 , yeah.

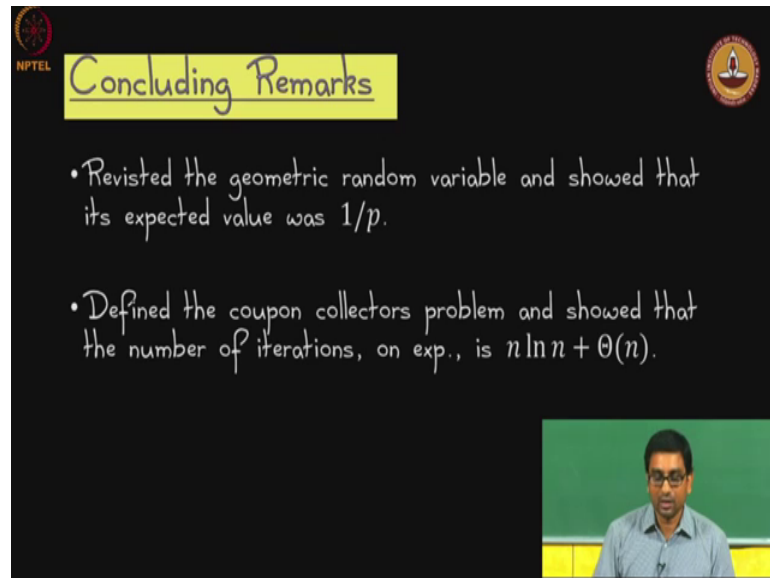
Student: (Refer Time: 18:52).

Ok what is what is the type of.

Student: T of X is θ of (Refer Time: 18:59) yeah. So, the statement (Refer Time: 19:01).

Oh ok, yes, yeah thank you. So, with that we conclude the proof of this theorem.

(Refer Slide Time: 19:19)



The slide features a dark background with a yellow title bar at the top containing the text "Concluding Remarks". In the top left corner, there is a small NPTEL logo, and in the top right corner, there is a circular logo of an institution. The main content consists of two bullet points written in white text:

- Revisited the geometric random variable and showed that its expected value was $1/p$.
- Defined the coupon collectors problem and showed that the number of iterations, on exp., is $n \ln n + \theta(n)$.

In the bottom right corner of the slide, there is a small rectangular inset showing a man with glasses and a light blue shirt speaking in front of a green chalkboard.

And so we can conclude this segment just to remind ourselves we revisited the geometric random variables shown that the expectation values 1 over p when the parameter is p . And we looked at the coupon collectors problem, and we show the expected number of times we need to buy the box of chocolates if you will is $n \ln n$ plus some θn .

Next segment we are going to again look at something interesting again algorithmic problem, finding the median over more generally the case selection problem.

Thank you.