

**Algorithms for Big Data**  
**Prof. John Ebenezer Augustine**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Madras**

**Lecture – 10**  
**Bernoulli, Binominal, and Geometric Distributions**

(Refer Slide Time: 00:15)

**Segment 7: Some Useful Distributions**

**Bernoulli Trial:** an experiment with exactly two outcomes.  
E.g., a toss of a coin.

Often, one outcome is called "success" and the other "failure."

$\Pr(\text{success}) = p$  and  $\Pr(\text{failure}) = (1 - p)$

**NPTEL** We can define an associated random variable:

The slide also features a small video inset of a man with glasses and a white shirt, and the NPTEL logo in the bottom left corner.

So, we have come to the final segment in this lecture. And this final segment, we are going to talk about some probability distributions.

Let us first start with a very simple experiment called Bernoulli trial, simple experiments in which there are exactly two outcomes. And there are many such experiments you can think of where the classic example being that of tossing a coin. In the context of Bernoulli trials, we normally talk about one of the outcomes is being a success, and the other one being a failure. In this Bernoulli trials are often used to model such failure situations. And the normal notation is that the probability of success is denoted  $p$ , and the probability here failure is denoted  $1 - p$ .

(Refer Slide Time: 01:14)

We can define an associated random variable:

$$X = \begin{cases} 1 & \text{if outcome is a success} \\ 0 & \text{if outcome is a failure} \end{cases}$$
$$E[X] = p \cdot 1 + (1 - p) \cdot 0 = p$$

The slide includes the NPTEL logo in the bottom left corner and a small video inset of a person in the bottom right corner.

Along with this Bernoulli trial, we can also associated Bernoulli random variable, and it was call as X for now; X will take the value of 1 typically if the outcome is a success and 0 otherwise. And now the question is what is the expectation of X. Well, there are two values that X can take X can either take the value 1 or it can take the value of 0. And each of these outcomes has to be, waited by the proper their respected probabilities and so the 1 occurs probability p because p is a success probability and 0 occurs probability 1 minus p, and therefore we can work it out to show that the expectation of X is p.

(Refer Slide Time: 02:10)

### Binomial Distribution

Consider  $n$  Bernoulli trials, each with success probability  $p$ .

Let  $X$  denote the numbr of successes. Then,  $X$  is said to follow the Binomial Distribution.

$$\Pr(X = j) = \binom{n}{j} p^j (1 - p)^{n-j}.$$

The slide includes the NPTEL logo in the bottom left corner and a small video inset of a person in the bottom right corner.

So, now that we have a handle on Bernoulli trials, when we run random variables and the expectation of such Bernoulli random variables. Let us move onto by the binomial distribution; in the binomial distribution, we have  $n$  Bernoulli trials each with success probability  $p$ .



(Refer Slide Time: 02:51)

Let  $X$  denote the number of successes. Then,  $X$  is said to follow the Binomial Distribution.

Diagram illustrating a sequence of trials:  $(1-p)$   $p$   $(1-p)$   $p$   $p$   $(1-p)$   $p$   $(1-p)$ . Below the diagram, arrows point to the probabilities:  $p^j$  and  $(1-p)^{n-j}$ .

$$\Pr(X = j) = \binom{n}{j} p^j (1-p)^{n-j}.$$

To compute the expectation of  $X$ , we can use linearity of expectation.

And here are random variable  $X$  denotes the number of successes that we see in those  $n$  Bernoulli trials such random variables said to follow the binomial distribution. It basically gives you the number of trial successes. So, what is the probability that this random variable  $X$  is equal to  $j$ ? Well what is that mean for  $X$  to equal  $j$  there been  $n$  trials,  $n$  Bernoulli trials, out of them  $j$  trials have come out successfully the rest of all been failures. So, there are been  $j$  successes, and the rest of all been failures. To this is nothing to say these are the particular  $j$  trials should be a success, there out of these  $n$  trials some  $j$  trials are been successes. How many ways can you actually choose  $j$  locations within the  $n$  trials that is  $n$  choose  $j$ .

And now that you specify that these are the success locations. We had to work out the probabilities, so this location is will be a successful probability  $p, p, p, p$ , so that all of them together would be  $p$  raise to the  $j$ , and all the other locations were to be failures.

So, all them will have to occur with probability  $1$  minus  $p, 1$  minus  $p, 1$  minus  $p$  and  $1$  minus  $p$ , so that is  $1$  minus  $p$  raise to the power  $n$  minus  $j$ , where since there are  $n$  location  $j$  have been success  $n$  minus  $j$  is the number of failures, and that is what you

have over there. And this is the probability of X taking the particular value j. Now we can use this to compute the expectation of X, so that is going to be very competitive, but this is where the linearity of expectation comes in very handy.

(Refer Slide Time: 05:32)

expectation.

Let  $X_1, X_2, \dots, X_i, \dots, X_n$  be the Bernoulli random variables associated with the Bernoulli trials. Then,

$$E[X] = E\left[\sum_{i=1}^n X_i\right]$$

Applying linearity of expectation, we get

$$E[X] = E[\sum_i X_i] = \sum_i E[X_i] = \sum_i p = np.$$

Linearity of exp.

So, in order to apply the linearity of expectation, we are going to assign random variable to each Bernoulli trials, these are sometimes call the indicator random variables, it is indicate whether a particular Bernoulli trials was a success or failure. So,  $X_1$  will be 1 if the first Bernoulli trials came out to success; otherwise be a 0 and so on. Hence, it is very easy to see that our binomial random variable X is simply the summation overall, i ranging from 1 to n of  $X_i$ (s).

Now, we can easily apply the linearity of expectation. So, what we want to understand is expectation of X and that is therefore going to be nothing but the expectation of the summation over i  $X_i$  or that is just basically obtain by saying of my expectation to both sides that is what we get. And now we apply the linearity of expectations and so the expectation goes inside and summation comes outside we get summation of i expectation of the individual  $X_i$ (s), but we already know that this expectation of an individual  $X_i$ (s) nothing but p that is what we have over here. And we are again summing and up over n different values of i in each case the p value is common. So, the total therefore, it is n times p which is the expectation of the random variable goes.

(Refer Slide Time: 07:37)

Consider repeating a Bernoulli trial until we get success. The number of trials  $X$  follows the geometric distribution.

$$\Pr(X = n) = (1 - p)^{n-1} p$$
$$E[X] = \frac{1}{p}$$

$p = \frac{1}{\binom{n}{2}}$

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

NPTEL

So, the third distribution kind of see now is called the geometric distribution. Here again Bernoulli trial play a part in the following manner. So, consider repeating the same Bernoulli trial several times (Refer Time: 07:51) until we get the first success. The number of trials that we need to perform is let say is denoted by  $X$ , and this  $X$  is said to follow the geometric distribution.

So, let us look at what is the probability that  $X$  will take on a particular values say  $n$ , what is this mean that in this means that the first  $n$  minus 1 trials, where all failures because have they had any one of them been a success then the experiment would have stopped. So, first  $n$  minus 1 trials where all failures and the  $n$ th trial was success, so that is why we have  $n$  minus 1 failure, and one the  $n$ th occurrence  $n$ th trial success.

With a little bit of work which we will skip for now we can show that the expectation of  $X$  is equal to  $1$  over  $p$ . This particular distribution is very useful in the analysis of in context randomize algorithms. And that suppose we have randomized algorithm to solve a particular problem and this randomize algorithm succeeds with some probability  $p$ , but it could also fail with some probability  $1$  minus  $p$ .

And in fact, we saw such randomize algorithm in the context of (Refer Time: 09:47) algorithm, the algorithm would succeed with probability  $p$  holds  $1$  over  $n$  choose  $2$ ; and with the probability  $1$  minus  $p$  it would fail. So, the question is how many times on expectation would we have to repeat the algorithm before its final is succeeds, well that

is nothing but  $1/p$ . In that case, we have  $X$  would be equal to  $1/p$  would be and choose to (Refer Time: 10:27) we repeat and choose two times on expectation we would see here success. So, you can see how this particular distribution can come in quite handy.

(Refer Slide Time: 10:44)

The Coupon Collector's Problem

We have  $n$  coupons.

A trial is the act of picking up a coupon (with replacement).

Here, the random variable  $X$  is defined as the number required to see all  $n$  coupons.

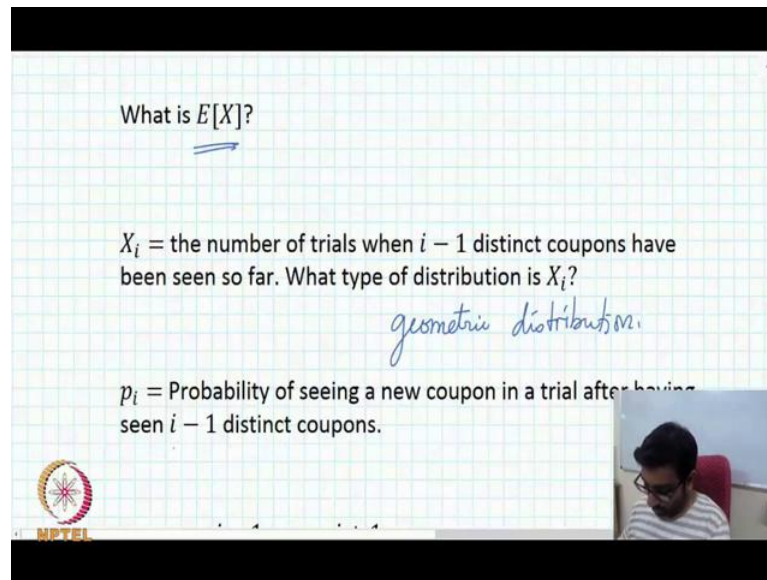
And now we will end with that that the last thing that we are going to see now is the coupon collector's problems that very interesting setting. So, we have  $n$  coupons, and you can think of these  $n$  coupons has been toys that are kept inside of one of those kinder joy candies that children buy all the time.

Each time buy a candy - the kinder joy chocolate you get a toy with it and. So, child is naturally interested in collecting all the toys that are available in this that he or she should possibly get. And each time the child buys the chocolate, it opens it up and only then can see can see whether the toys something that at it is already seen before or whether it is a new toy. And the question as in this coupon got just problem is how many chocolates do you need to buy before the child gets to collect all the  $n$  different toys that are get inside this chocolate casings.

Let us look at this problem a bit more carefully. So, we have  $n$  coupons the trial is the act of picking of a coupon with replacements. So, slight buying a chocolate and there is so many chocolates that buying one chocolates does not really effect the functionality. So, you for simplicity we just assume that these chocolates or these toys are drawn with

replacement. So, so these coupons are, you look at the coupon, and then you see whether a coupon that you have already seen before, if you seen it before well that is the wasted trail, if you never seen it before well that means, you seen a new coupon. You put coupon that and then and then you again randomly pick one more and so on.

(Refer Slide Time: 13:13)



And so here the random variable  $X$  is defined as the number of trials required to see all the  $n$  coupons. So, this is called the coupon collector's problems, and what is this expectation of this random variables. Well, again we are going to use linearity of expectation. And for this, we again have to sort of decompose the random variable  $X$  into component  $X_i$ (s) in the appropriate way, here is how we going to do it.  $X_i$  is going to be the random variable that denotes the number of trials when  $i$  minus 1 distinct coupon have been seen so far.

What type of distribution is  $X_i$ ? Well, if you have seen  $i$  minus 1 coupon so far well. So, you pick a coupon, if it is a new coupon that you never seen before it is a success that you have done,  $X_i$  has is already that particular face is already over. But if it is a coupon that you have already seen before then that is a wasted trial, you have to repeat again. So, if you think about this is nothing but the geometric distribution. And here the probability of success  $p_i$  is the probability of seeing a new coupon in a trial after having seen  $i$  minus 1 distinct coupons.

(Refer Slide Time: 15:16)

$$p_i = 1 - \frac{i-1}{n} = \frac{n-i+1}{n}$$
$$E[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$$
$$E[X] = E\left[\sum X_i\right] = \sum E[X_i] = \sum \frac{n}{n-i+1}$$

So, what is this  $p_i$  well that is a, first look at this is the probability that you will actually see a coupon that you have already seen before, there are  $i$  minus 1 coupon that you have already seen before and there are total of  $n$  coupons. So, this is a probability of seeing a coupon that you have already seen before and 1 minus of that quantity is the probability that you have going to see and you that is  $p_i$ . And it works out to be  $n$  minus  $i$  plus 1 that (Refer Time: 15:51)  $n$ . And since  $X_i$  is following the geometric distribution  $E$  of  $X_i$  is 1 by  $p_i$  which is equal to  $n$  divided by  $n$  minus  $i$  plus 1.

(Refer Slide Time: 16:08)

$$E[X] = E\left[\sum X_i\right] = \sum E[X_i] = \sum \frac{n}{n-i+1}$$
$$= n \sum \frac{1}{n-i+1} = n \sum \frac{1}{i} = n H_n \in \Theta(n \log n).$$

*n # Harmonic number*



So, now we can apply the linearity of expectation  $E$  of  $X$  is equal to  $E$  of the summation over  $i$   $X_i$  and then you apply the linearity of expectation you get that you apply the formula for  $E$  of  $X_i$  that we have seen here. And this  $n$  can be taken out of the summation you get a summation over  $i$  one of the  $i, n - i + 1$ .

Just rearranging the summation terms, you get summation over  $i$   $1/i$  and this is nothing but the  $n$ th harmonic number. So, you will get  $n$  times  $H_n$  this is the  $n$ th harmonic number, and we know that this  $H_n$  is going to be roughly  $\log n$ . So, this quantity is  $\Theta(n \log n)$ .