

Essentials of Data Science with R Software – 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture – 36
Geometric Distribution

Hello friends, welcome to the course Essentials of Data Science with R Software – 1; in which we are trying to understand the basic fundamental topics of probability theory and statistical inference. So, up till now you can see we have considered a couple of probability mass function, discrete uniform, Bernoulli, degenerate, binomial, Poisson.

And you have observed that these are special types of functions, which are trying to describe the behavior of particular type of phenomena or a process. So, now the advantage is that if you can identify phenomena which is resembling with any of this probability function; you can very easily create a probability model. And that is one of the basic objectives in the data science.

So, now similarly you can believe on me that there is a long list of such probability mass functions. And they have got some application; they are trying to describe particular type of phenomena and so on. So, from that list today we are going to choose one new probability distribution, which is geometric distribution.

So, now the first question comes that as soon as I take the name of a new probability mass function or probability distribution, where it is going to use? How it is going to use? And what type of events this can model? What type of probabilities of events can be modeled using the geometric distribution?

So, let us try to take an example to understand this thing. Suppose box has got two types of balls: black and white; and suppose you want to draw a white ball. So, what you try to do? That you try to shake the box, close your eyes and put your hand inside the box and take out the ball. You wanted to take out the white ball, but this time it comes out to be a black ball; that you do not want. So, what you try to do? You will once again try to draw another ball; and suppose this balls once again comes out to be black is still you do not want. So, you try to make the third draw and the third draw suppose you get a white ball.

So, that means now you are happy, you are successful; so, if you try to see that in the third draw, you got a success. And in the remaining two draws, they were sort of fail not really a failure. But what you really wanted? You wanted a white ball; and then you were getting a black ball in that sense it is a failure. So, now we go back to the point where we started the experiment. Now, I ask you same question in a different way. This box has got white and black balls; I ask you to take out one ball at a time. And try to inform me in which of the draw, you will get a white ball?

So, now that is possible that in the first draw itself you can get the white ball; or in the second draw, you will get the white ball. That means the first draw was black ball or in the third draw, you will get the white ball; so that in the first two draw they were black balls. Similarly, if I say suppose you get the white ball in the seventh draw; so that means before that up to sixth draw, you got only the black balls.

So, what you are trying to observe here? What is really happening that you are getting a success at a particular draw? And you want to know the probability that what is the probability that we will get the success at this particular draw? Or, we want to know that, what the probability that in the seventh draw, we will get a white ball.

So, in this case, you can see that each of the event is just like a Bernoulli trial; there are only two outcomes, either you are getting a success or a failure. Success and failure means you are getting a white ball or a black ball. So, in these types of situation, where you want to compute the probability of getting a success; success at the r th trial, say for example.

Then you try to make use of geometric distribution and similarly there can be many many situations, where you can compute these types of probabilities. You will actually see that many such situations come in practice, where you would like to use the geometric distributions to directly compute the probabilities and other properties.

So, with this example let us try to comeback to our slides once again and try to understand the different aspects of geometric distribution. This lecture I will try to concentrate on the theoretical properties, and in the next lecture I will try to show you how you can do the things in the R software. So, let us begin.

(Refer Slide Time: 06:05)

Geometric Distribution:
Consider a random experiment that is closely related to the one used in the definition of a binomial distribution.

Again, assume a series of Bernoulli trials (independent trials with constant probability p of a success on each trial).

However, instead of a fixed number of trials, trials are conducted until a success is obtained.

Let the random variable X denote the number of trials until the first success.

So, now say consider a random experiment that is closely related to the one used in the definition of a binomial distribution. You know that you had considered the Bernoulli trials. So, assume a series of Bernoulli trials that means independent trials with constant probability p of a success on each trial. So, however instead of a fixed number of trial, trials are conducted until a success is obtained. So, now we do not know whether the success is going to obtain in the first draw, or in the second draw, or in the third draw and so on. So, in that sense the random variable X is indicating the number of trials until the first success; and this number is going to be a random variable.

(Refer Slide Time: 06:59)

Geometric Distribution:
Consider a situation of drawing of lottery ticket in a draw.
Every draw results in "win" or "lose".

Suppose we want to know how many lottery tickets are needed to buy until we win for the first time.

Suppose we are interested in determining how many independent Bernoulli trials are needed until the event of interest occurs for the first time.

The outcomes "win" or "lose" are Bernoulli trials.

So, for example, consider a situation where there is a lottery ticket, and there are two possible options that we can lose the ticket or we can win the ticket. So every draw of the lottery results into win or lose; and suppose we want to know that how many lottery tickets are needed to buy, until we win the lottery for the first time. So suppose we are interested in determining how many independent Bernoulli trials are needed, until the event of interest occurs for the first time. That is the same thing, but in our statistical language. The outcomes which I have just considered win or lose; they are essentially the outcomes of a Bernoulli trial.

(Refer Slide Time: 07:46)

Geometric Distribution:
Similarly, suppose drugs are being tried in a clinical trial to treat the disease successfully.
We want to know how many different drugs to try to successfully tackle the disease, etc.
The geometric distribution can be used to determine the probability that the event of interest happens at the k^{th} trial for the first time.

Similarly, you can consider one more example. Suppose drugs are being tried on a in a clinical trial to treat the disease successfully; and we want to know that how many different drugs to try to successfully tackle that disease. That will also have a different type of situation in biostatistics. So, under this type of situation, we can use the geometric distribution. And we can compute the probability that the event of interest happens at the say k^{th} trial, for the first time; k can take different integer values.

(Refer Slide Time: 08:30)

Geometric Distribution:
 A discrete random variable X is said to follow a geometric distribution with parameter p if its PMF is given by

$$P(X = k) = p(1 - p)^{k-1}, k = 1, 2, 3, \dots$$

The mean (expectation) and variance are given by

$$E(X) = \frac{1}{p} - 1$$

$$Var(X) = \frac{1}{p} \left(\frac{1-p}{p} \right)$$

So, now based on this we can define here the geometric distribution. So, a discrete random variable X is said follow a geometric distribution with parameter p , if its probability mass function is given by like this $P(X = k) = p(1 - p)^{k-1}, k = 1, 2, 3, \dots$

. And in case if you try to find out the mean and variance of this random variable X following the geometric distribution.

Then the mean comes out to be here $\left(\frac{1}{p} - 1\right)$; and the variance comes out to be here $\frac{1}{p}$. And inside the parenthesis that is multiplied by $\left(\frac{1}{p} - 1\right)$. So, this is something like $\frac{1}{p} \left(\frac{1}{p} - 1\right)$; so that is the variance. Well, means I am not giving you the proof, but but if you wish, you can do a very simple algebra to find out these expressions.

(Refer Slide Time: 09:30)

Geometric Distribution:

Example:

The probability that a machine produces a faulty transistor is 0.1.

Assume the production of transistors are independent events, and let the random variable X denote the number of transistors produced until the first error.

Then, $P(X = 5)$ is the probability that the first four transistors are produced correctly and the fifth transistor has error.

This event can be denoted as $\{GGGGD\}$, where G denotes a good transistor and D denotes a defective transistor.

So, let us try to take one more example to understand now. Suppose the probability that a machine produces a faulty transistor is 0.1. And assume the production of transistors are independent events, and let the random variable X denote the number of transistors produced until the first error occurs.

So, for example, if you want to find out the probability that X equal to 5; this means this is the probability that the first 4 transistors are produced correctly, and the fifth transistor has got some error. So, now in case if you want to denote such an event, we can denote by say $GGGG$ and D . So, G means here good and D means here defective. So, this is a sequence of events in which it is the, it is occurring.

(Refer Slide Time: 10:28)

Geometric Distribution:
Example:

Because the production of transistors are independent and the probability of a correct transistor is 0.9

$$P(X = 5) = P(GGGGD) = P(G) \times P(G) \times P(G) \times P(G) \times P(D)$$
$$= 0.9^4 \cdot 0.1 = 0.066$$

Handwritten notes: 1st, 2nd, 3rd, 4th, 5th transistor

And now you know the productions of these transistors are independent and the probability of a correct transistor is $1 - 0.1$, which is 0.9 . So, that means if you want to know the probability of X equal to 5 ; that means you are getting the first defective transistor in the fifth draw. That means first 4 draws are independent and they have the probability, probability of G into probability of G into probability of G into probability of G ; where this is for the first transistors. The second probability is for the second transistor, third probability is for the third transistor, and fourth probability is the fourth transistors probability.

And fifth is the probability of the fifth transistor that we want. So, if you try to simply solve it, this will come out to be 0.066 . So, there are almost 6 percent chances or 6.6 percent chances approximately; that the fifth predication of the transistor will give us an error.

(Refer Slide Time: 11:42)

Geometric Distribution:
Example: Suppose a coin is tossed until "head" is obtained for the first time. The probability of getting a head is $p = 0.5$ for each toss.

$$P(X = 1) = 0.5$$

$$P(X = 2) = 0.5(1 - 0.5) = 0.25$$

$$P(X = 3) = 0.5(1 - 0.5)^2 = 0.125$$

$$P(X = 4) = 0.5(1 - 0.5)^3 = 0.0625$$

.....

The mean and variance are

$$E(X) = \frac{1}{0.5} = 2$$

$$Var(X) = \frac{1}{0.5} \left(\frac{1}{0.5} - 1 \right) = 2$$

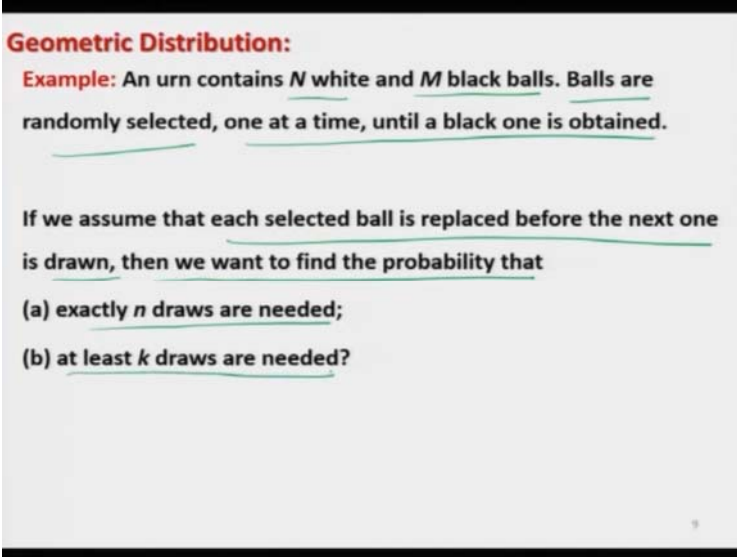
Similarly, if I to take one more example here that suppose, a coin is tossed until the head is obtained for the first time. So, the probability of getting a head or a tail in the toss of a fair coin is 0.5 that we know; so, and this is a Bernoulli trial obviously. So, probability that X equal to 1 , that you get the success in the first trial is simply 0.5 , half.

And if you want to know the probability of success in the second trial; this means with the probability at a first trial 0.5 ; and the probability of the second trial, which is $1 - p$, which $1 - 0.5$.

And similarly, if you want to know the probability of X equal to 3; that means this probability is going to be $0.5 (1 - 0.5)^2$.

And similarly, for X equal to 4, you can obtain the similar probability; so, this is what we mean by this geometric distribution. And in case if you try to find out here the mean and variance of this X ; they will come out to be $1/0.5$, which is $1/p$. So, this is a 2 and variance of X is $1/0.5$ into $1/0.5 - 1$, which is once again 2.

(Refer Slide Time: 12:59)



Geometric Distribution:
Example: An urn contains N white and M black balls. Balls are randomly selected, one at a time, until a black one is obtained.
If we assume that each selected ball is replaced before the next one is drawn, then we want to find the probability that
(a) exactly n draws are needed;
(b) at least k draws are needed?

Now, let me try to take one more example to make you understand; suppose, an urn contains N white and M black balls. The balls are randomly selected one at a time until a black one is obtained. So, if you assume that each selected ball is replaced before the next one is drawn; then we want to find the probability that exactly n draws are needed, and at least k draws are needed.

(Refer Slide Time: 13:34)

Geometric Distribution:

Let X : Number of draws needed to select a black ball, then X follows the geometric distribution with $p = \frac{M}{M+N}$

(a) Hence probability of exactly n draws is obtained as.

$$P(X = n) = \left(\frac{N}{M+N}\right)^{n-1} \left(\frac{M}{M+N}\right) = \frac{MN^{n-1}}{(M+N)^n}$$

So, how to solve this type of situation? So, definitely in this case the the probability distribution of geometric distribution can be used without any problem. So, let the random variable X indicate the number of draws needed to select a black ball. So, then obviously the black ball can be selected in the first draw, or in the second draw, or and so on.

So, this will follow a geometric distribution and the probability will be $M/(M + N)$. So, the probability of exactly a small n number of draws can be obtained by putting probability of X equal to n ; which is simply here $\left(\frac{N}{M+N}\right)^{n-1} \left(\frac{M}{M+N}\right)$. And you can solve it; you will get such a probability.

(Refer Slide Time: 14:29)

Geometric Distribution:

(b) Hence probability of at least k draws are needed, is obtained as:

Since the probability that at least k trials are necessary to obtain a success is equal to the probability that the first $(k - 1)$ trials are all failures. Thus for a geometric random variable

$$P(X \geq k) = (1 - p)^{k-1}$$

with $p = \frac{M}{M+N}$.

Alternatively

$$P(X \geq k) = \left(\frac{M}{M+N}\right) \sum_{n=k}^{\infty} \left(\frac{N}{M+N}\right)^{n-1} = \frac{\left(\frac{M}{M+N}\right) \left(\frac{N}{M+N}\right)^{k-1}}{\left(1 - \frac{N}{M+N}\right)} = \left(\frac{N}{M+N}\right)^{k-1}$$

So, this is not difficult at all, you can see. And once you obtain this thing, the probability of at least k draws is needed, is obtained as follows. Because, the probability that at least k trials are necessary to obtain a success is equal to the probability, that the first $k - 1$ trials are all failures; only then, at the k trials, you will get the success. So, then in the terms of geometric random variable, we want to find out here the probability that X is greater than or equal to k , which is simply $(1 - p)^{k-1}$; where, p is given by $M/(M + N)$.

And alternative also means you can compute this probability by the by this expression also, that probability of X greater than equal to k . That is equal to $\left(\frac{M}{M+N}\right) \sum_{n=k}^{\infty} \left(\frac{N}{M+N}\right)^{n-1}$. And if you simply try to solve it, you get here the same thing what you have obtained.

(Refer Slide Time: 15:36)

Geometric Distribution: Lack of Memory Property

A geometric random variable is defined as the number of trials until the first success.

However, because the trials are independent, the count of the number of trials until the next success can be started at any trial without changing the probability distribution of the random variable.

For example, in the production of transistors example, if 100 transistors are produced without any defect and suppose the 106th transistor is defective.

So the first error occurs on 106th transistor.

In the next lecture, I will try to take an example in the R and I will try to illustrate these things. Now, this geometric distribution has a very specific property, which is called as lack of memory. This lack of memory property is having in the geometric distribution and we will try to see later on. And that it is also there in one of the continuous probability density functions; that are for the continuous random variable, so that we will try to see later on. So, now the question is what is this lack of memory property? The understanding of this concept is very important. And I personally believe that this is really going to help you in the data science.

So, we know that a geometric random variable is defined as the number of trial until the first success come. Now, because the trials are independent for the count of the number of trials, until the next success can be started at any trial without changing the probability distribution of the random variable. This is the basic idea behind this lack of memory property. So, let me try to explain you this thing through an through a very simple example; that for example in the example of production of the transistors. Suppose, if 100 transistors are produced without any defect and suppose the 106th transistor is defective. So, that means the first error is occurring at 106th transistor.

(Refer Slide Time: 17:27)

Geometric Distribution: Lack of Memory Property

Thus the probability that the next six outcomes after 100th transistor, i.e., for the event denoted as GGGGGD is

$$P(X = 6) = P(GGGGGD) = P(G) \times P(G) \times P(G) \times P(G) \times P(G) \times P(D)$$
$$= 0.9^5 \cdot 0.1 = 0.059$$

This probability is identical to the probability that the initial error occurs on 6th transistor.

The implication of using a geometric model is that the system presumably will not wear out.

The probability of an error remains constant for all transistor.

In this sense, the geometric distribution is said to lack any memory.

So, the probability that the next six outcome after the 100th transistor; that is for the event that is indicated by say here GGGGGD is simply probability that X equal to 6; which is here. The trials are independent; so that we already have obtained this will be something like 0.059. And this probability is identical to the probability that the initial error occurs on the 6th transistor. So, now what is the meaning? The implication of using a geometric model is that the system presumably will not wear out. The probability of an error remains constant for for all the transistors.

And in this sense the geometric distribution is said to have the lack of memory property; that it is said to lack any memory. So, now we come to an end to this lecture, and I have given you the basic fundamental definitions related to the geometric distribution. And in the next lecture I will try to show you the application of geometric distribution in the R software.

But, once again I will request you the same that try to look into assignment, try to look into books; and try to solve some examples, and try to read from the books also. Definitely, the books cannot replace these lectures. I have here limited time, in which I am trying to manage; but with the books, you have lot of time. Try to study, try to practice and I will see you in the next lecture. Till then good bye.