Statistics for Data Science -1

Lecture: Hypergeometric Distribution

Usha Mohan

Indian Institute of Technology Madras

Learning objectives

Learning objectives

1. Derive the formula for the probability mass function for Hypergeometric distribution.

Learning objectives

1. Derive the formula for the probability mass function for Hypergeometric distribution.

Learning objectives

1. Derive the formula for the probability mass function for Hypergeometric distribution.
2. Expectation and variance of the Hypergeometric distribution.

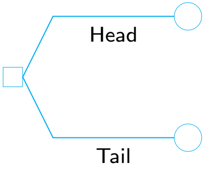
Learning objectives

1. Derive the formula for the probability mass function for Hypergeometric distribution.
2. Expectation and variance of the Hypergeometric distribution.
3. To understand situations that can be modeled as a Hypergeometric distribution.

Limitations of Binomial distribution

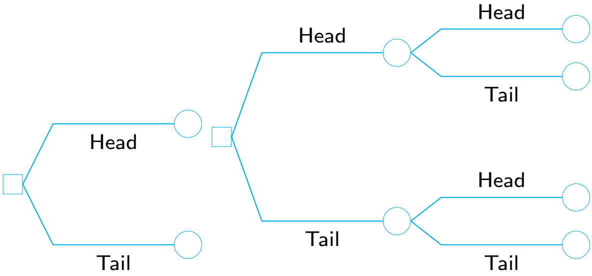
Limitations of Binomial distribution

- Suppose we are interested in finding the probability of success in n trials, where the trials are independent.



Limitations of Binomial distribution

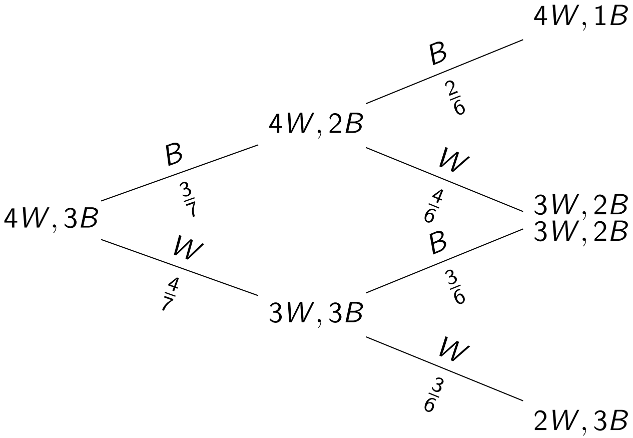
- Suppose we are interested in finding the probability of success in n trials, where the trials are independent.



Limitations of Binomial distribution

Limitations of Binomial distribution

- Suppose trials are not independent and the probability of "success" is not the same for all trials.



Introduction

For the hypergeometric to work,

- The population must be dividable into two and only two independent subsets (black balls and white balls in our example).
- The experiment must have changing probabilities of success with each experiment (the fact that balls are not replaced after the draw in our example makes this true in this case).
- Another way to say this is that you sample without replacement and therefore each draw is not independent.

The Hypergeometric distribution

- A discrete random variable (RV) that is characterized by:
    - A fixed number of trials.
    - The probability of success is not the same from trial to trial.
- We sample from two groups of items when we are interested in only one group.
- X is defined as the number of successes out of the total number of items chosen.

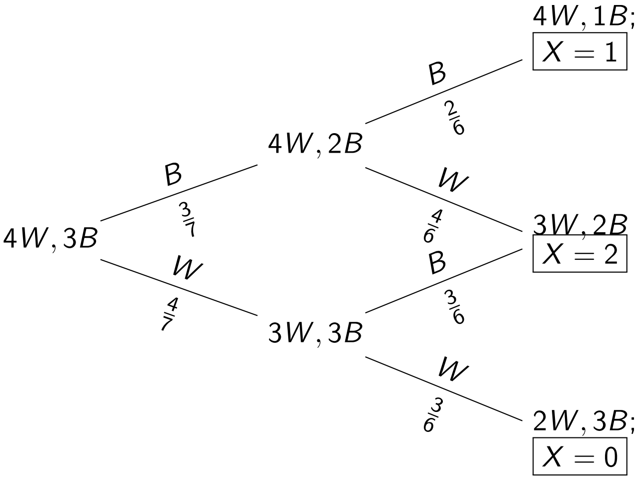
Understanding the Hypergeometric distribution

If we randomly select n items without replacement from a set of N items of which: m of the items are of one type and $N - m$ of the items are of a second type

$$\overbrace{\underbrace{\boxed{\text{B}}\ \boxed{\text{B}}\ldots\boxed{\text{B}}}_{m}\ \underbrace{\boxed{\text{W}}\ \boxed{\text{W}}\ \ldots\boxed{\text{W}}}_{N-m}}^{N} \curvearrowright \overbrace{\underbrace{\boxed{\text{B}}\ \ldots\boxed{\text{B}}}_{x}\ \underbrace{\boxed{\text{W}}\ \ldots\boxed{\text{W}}}_{n-x}}^{n}$$

Let X be the number of items of type 1, then the probability mass function of the discrete random variable, X, is called the hypergeometric distribution and is of the form:

$$P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}; x = 0, 1, \ldots, n$$


$4W, 1B;$
$\boxed{X = 1}$
B
$\frac{2}{6}$
$4W, 2B$
B
$\frac{3}{7}$
W
$\frac{4}{6}$
$3W, 2B$
$\boxed{X = 2}$
$4W, 3B$
W
$\frac{4}{7}$
B
$\frac{3}{6}$
$3W, 3B$
W
$\frac{3}{6}$
$2W, 3B;$
$\boxed{X = 0}$

Examples: Choosing balls without replacement

A bag consists of 7 balls of which 4 are white and 3 are black. A student randomly samples two balls without replacement. Let X be the number of black balls selected.

- Here, $N = 7, n = 2, m = 3$
- X takes values: $0, 1, 2$
- $P(X = i) = \dfrac{\binom{3}{i}\binom{4}{2-i}}{\binom{7}{2}}; i = 0, 1, 2$
- The pmf

i	0	1	2
P(X=i)	$\frac{12}{42}$	$\frac{24}{42}$	$\frac{6}{42}$

Examples: Choosing balls without replacement

A bag consists of 50 balls of which 30 are white and 20 are blue. A student randomly samples five balls without replacement. Let X be the number of blue balls selected.

- Here, $N = 50, n = 5, m = 20$
- X takes values: $0, 1, 2, 3, 4, 5$
- $P(X = i) = \dfrac{\dbinom{20}{i}\dbinom{30}{5-i}}{\dbinom{50}{5}}; i = 0, 1, 2, 3, 4, 5$
- The pmf

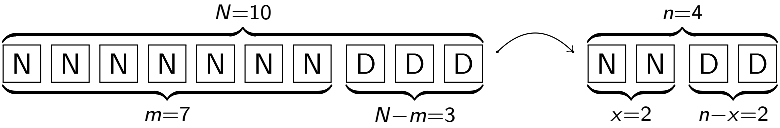
i	0	1	2	3	4	5
P(X=i)	0.07	0.26	0.36	0.23	0.07	0.01

Example: Voters

- Assume there are 150 female voters and 250 male voters in a particular locality. If a group of twenty five voters is selected at random, then the probability that ten of the selected voters would be female can be calculated with the help of hypergeometric probability distribution.
- In this case: $N = 400, n = 25, m = 150$
- $$P(X = 10) = \frac{\binom{150}{10}\binom{250}{15}}{\binom{400}{25}}$$

Example: Defectives-1

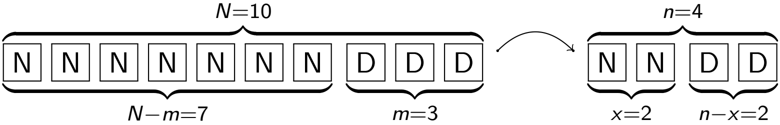
- In a batch of 10 computer parts it is known that there are three defective parts. Four of the parts are selected at random to be tested. Define the random variable X to be the number of working (non defective) computer parts selected



- This is a Hypergeometric distribution with $N = 10, n = 4, m = 7$
- The pmf $P(X = x) = \dfrac{\binom{7}{x}\binom{3}{4-x}}{\binom{10}{4}}; x = 0, 1, 2, 3, 4$

Example: Defectives-2

- In a batch of 10 computer parts it is known that there are three defective parts. Four of the parts are selected at random to be tested. Define the random variable X to be the number of defective computer parts selected



- This is a Hypergeometric distribution with $N = 10, n = 4, m = 3$
- The pmf $P(X = x) = \dfrac{\binom{3}{x}\binom{7}{4-x}}{\binom{10}{4}}; x = 0, 1, 2, 3$

Example: Sampling from a deck of cards

- Take a deck of 52 cards. Draw five cards from the deck. Let the random variable X denote the number of aces in the random sample of five cards. What is the probability distribution of X?
- This is a Hypergeometric distribution with $N = 52, n = 5, m = 4$
- The pmf is given by

$$P(X = x) = \frac{\binom{4}{x}\binom{48}{5-x}}{\binom{52}{5}}; x = 0, 1, 2, 3, 4$$

Hypergeometric distribution

If we randomly select n items without replacement from a set of N items of which: m of the items are of one type and $N - m$ of the items are of a second type

The probability mass function of the discrete random variable is called the hypergeometric distribution and is of the form:

$$P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}; x \leq n, x \leq m, n - x \leq N - m$$

Section summary

- Understanding the hypergeometric distribution.
- Obtaining the probability mass function of the distribution.