1. Introduction

Welcome to the series of e-learning modules on Hypergeometric Distribution. In this module we are going to study about hypergeometric distribution, its mean and variance, factorial moments, approximation to binomial distribution, recurrence relation for the probabilities and application.

By the end of this session, you will be able to explain:

- About hypergeometric distribution
- Mean
- Variance
- Factorial moments
- Approximation to binomial distribution
- Recurrence relation for probabilities
- Application

Introduction

When the population is finite and the sampling is done without replacement, so that the events are stochastically dependent, although random, we obtain hypergeometric distribution. Hypergeometric distribution was first used in the year 1936.

Consider an urn with N balls, M of which are white and N minus M are red. Suppose that we draw a sample of n balls at random without replacement from the urn, then the probability of getting k white balls out of n where k is less than n, is
M c k into N minus M c n minus k divided by N c n.

Definition

A random variable X is said to follow hypergeometric distribution with parameters N, M and n if its probability mass function is given by
P of x is equal to k is equal to M c k into N minus M c n minus k divided by N c n, k take values zero 1, 2, etc minimum of n and M.

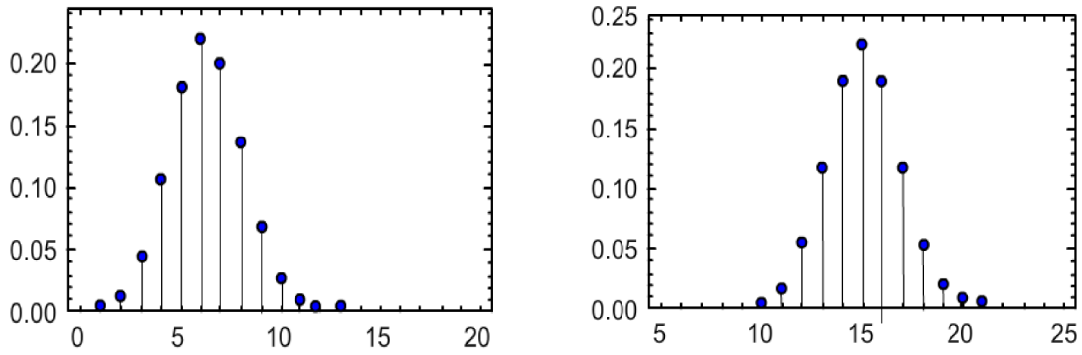
As it can be shown that,
summation over k is equal to zero to n, M c K into N minus M c n minus k divided by N c n is equal to 1, the assignment of probabilities is permissible.

The basic characteristics of hypergeometric distribution is,

- It models the success of the process of sampling without replacement from a finite population
- It differs from the binomial distribution only in the fact that the population is finite and the sampling from the population is without replacement
- Trials are dependent

The graph of probability mass function is shown below.

Figure 1



Let us now have a look at a few examples of hypergeometric distribution
Say, Choosing a team of 8 from a group of 10 boys and 7 girls.

Choosing a committee of five from the legislature consisting of 52 Democrats and 48 Republicans.

Choosing exactly two marbles of each colour from 5 black, 10 white and 15 red marbles.

Choosing at least one defective chip when a batch of 100 computer chips containing 10 defective chips and 5 chips are chosen at random.

Suppose, a pond contains 1000 fish and 200 are tagged and a sample of size 20 is taken then getting 5 tagged fishes from the sample.

2. Mean and Variance of Hypergeometric Distribution

Now let us find the mean of the distribution.
Mean is given by, expectation of x is equal to summation over k from zero to n, k into p of x is equal to k
Is equal to summation over k, k into M c k into N minus M c n minus k divided by N c n.
Is equal to M divided by N c n, summation over k is equal to 1 to n M minus 1 c k minus 1 into N minus M c n minus k.

Is equal to M divided by N c n into summation over x from zero to m A c x into N minus A minus 1 c m minus x, where x is equal to k minus 1, m is equal to n minus 1 and M minus 1 is equal to A.
Is equal to M divided by N c n into N minus 1 c m
Is equal to M divided by N c n into N minus 1 c n minus 1,
Which, is equal to n into M by N.
Therefore mean of the distribution is n into M by n.

Now let us find the variance of the distribution. Here, first we find expectation of x square, which can be written as, expectation of x into x minus 1 plus x.
So, considering expectation of x into x minus 1
Is equal to summation over k is equal to zero to n, k into k minus 1 into p of x is equal to k
Is equal to summation over k, k into k minus 1 into M c k into N minus M c n minus k divided by N c n
Is equal to M into M minus 1 divided by N c n into summation over k from 2 to n, M minus 2 c k minus 2 into N minus M c n minus k

Is equal to M into M minus 1 divided by N c n into N minus 2 c n minus 2
Is equal to M into M minus 1 into n into n minus 1 divided by N into N minus 1.

Therefore expectation of x square is equal to expectation of x into x minus 1 plus expectation of x
Is equal to M into M minus 1 into n into n minus 1 divided by N into N minus 1 plus n into M by n.
Hence variance of x is equal to expectation of x square plus expectation of x the whole square.
By substituting the values of expectations, we get,
V of x is equal to M into M minus 1 into n into n minus 1 divided by N into N minus 1 plus n into M by N plus n into M by N the whole square.

By taking n into M by N outside, we get
n into M divided by N into M minus 1 into n minus 1 divided by N minus 1, plus 1 plus n into M by N
taking common denominator and simplifying, we get,
n into M into N minus M into N minus n divided by N square into N minus 1.

3. Limiting Form of Hypergeometric Distribution

Now let us find the limiting form of hypergeometric distribution.
Hypergeometric distribution tends to binomial distribution as N tends to infinity and M by N tends to p.

Consider p of x is equal to k
Is equal to M c K into N minus M c n minus k divided by N c n

Writing using factorials we get,
M factorial divided by k factorial into M minus k factorial, into N minus M factorial divided by n minus k factorial into N minus M minus n plus k factorial, into n factorial, into N minus n factorial divided by N factorial.

Expanding numerators of first and second term to cancel the common terms with the denominator and expanding denominator of third term and cancelling the common terms with numerator, we get,
M into M minus 1 into M minus 2 into etc., M minus k plus 1 divided by k factorial, into
N minus M into N minus M minus 1 into etc., N minus M minus k plus 1 divided by n minus k factorial, into
n factorial divided by N into N minus 1 into N minus 2 into etc., N minus n plus 1.

By combining the factorial terms and rearranging to apply the limit, we get
n factorial by k factorial into n minus k factorial into, M by N into M by N minus 1 by N into M by N minus 2 by N into etc., M by N minus k minus 1 by N into,
1 minus M by N into 1 minus M by N minus 1 by N into etc., 1 minus M by N minus n minus k minus 1 by N divided by,
1 minus 1 by N into 1 minus 2 by N into etc., 1 minus n minus 1 by N.

Factorial terms we can write as n c k. Therefore p of x is equal to
n c k into, M by N into M by N minus 1 by N into M by N minus 2 by N into etc., M by N minus k minus 1 by N into,
1 minus M by N into 1 minus M by N minus 1 by N into etc., 1 minus M by N minus n minus k minus 1 by N divided by,
1 minus 1 by N into 1 minus 2 by N into etc., 1 minus n minus 1 by N.

Proceeding to the limit as N tends to infinity and M by N tends to p we get,
Limit n tends to infinity p of x is equal to

n c k into, limit n tends to infinity, M by N, into M by N minus 1 by N, into M by N minus 2 by N, into etc., M by N minus k minus 1 by N into,
1 minus M by N into 1 minus M by N minus 1 by N into etc., 1 minus M by N minus n minus k minus 1 by N divided by,
1 minus 1 by N into 1 minus 2 by N into etc., 1 minus n minus 1 by N.

Is equal to n c k into p into p into etc., into p, k times, into 1 minus p into 1 minus p into etc., into 1 minus p, n minus k times
is equal to n c k into p power k into 1 minus p power n minus k, which is the probability mass function of binomial distribution with parameters n and k. Hence limiting distribution of hypergeometric distribution is binomial distribution.

4. Recurrence Relation of Hypergeometric Distribution

Recurrence relation for the probabilities of the hypergeometric distribution can be found as follows.
We know that p of x is equal to k, is equal to M c k into N minus M c n minus k divided by N c n
Consider p of x is equal to k plus 1, is equal to M c k plus 1 into N minus M c n minus k minus 1 divided by N c n.

Now consider the ratio of p of x is equal to k plus 1 divided by p of x is equal to k. we get,

M c k plus 1 into N minus M c n minus k minus 1 divided by N c n whole divided by, M c k into N minus M c n minus k divided by N c n

Writing in terms of factorials and expanding and then cancelling the common terms in numerator and denominator and then simplifying we get,
n minus k into M minus k, divided by k plus 1 into N minus M minus n plus k plus 1.

Hence the recurrence relation for the probabilities of hypergeometric distribution is given by,
P of X is equal to k plus 1 is equal to n minus k into M minus k whole divided by k plus 1 into N minus M minus n plus k plus 1 into p of x is equal to k, for k is equal to 1, 2, 3 etc.

Now let us find the factorial moments of hypergeometric distribution
The R^{th} factorial moment is given by,
Expectation of X power r is equal to summation over k is equal to r to n, k power r into p of x is equal to k
Is equal to summation over k, k power r into M c k into N minus M c n minus k divided by N c n
Expanding k power r and M c k and then simplifying we get,
summation over k, M power r into M minus r c k minus r into N minus M c n minus k divided by N c n

By substituting j with k minus r and doing the adjustments, we get
M power r into summation over j from zero to n minus r, M minus r c j into N minus r minus of M minus r c n minus r minus j divided by N c n.

Now simplifying the last term, we get,

M power r into n power r divided by N power r into summation over j,
M minus r c j into N minus r minus of M minus r c n minus r minus j divided by N minus r c n minus r.

Is equal to M power r into n power r divided by N power r into one
Is equal to M power r into n power r divided by N power r.

Therefore expectation of x power r is equal to M power r into n power r into N power r, where x power r is equal to x into x minus 1 into etc., x minus r plus one.

5. Application of Hypergeometric Distribution

Now we will consider the hypergeometric model to estimate the number of fish in a lake.

Let us suppose that in a lake there are 'N' fishes, 'N' being unknown. The task here is to estimate 'N'. A catch of 'r' fish, all at the same time, is made and these fishes are returned alive into the lake after marking each with a red spot.

After a reasonable period of time, during which the 'marked' fishes are assumed to have distributed themselves 'at random' in the lake, another catch of 's' fishes, again all at once, is made. Here r and s are regarded as fixed predetermined constants. Among these s fishes caught, there will be, say, x marked fishes, where, X is a random variable following discrete probability function given by hypergeometric model.

p of N is equal to r c x into N minus r c s minus x divided by N c s, where x is an integer such that maximum of zero and s minus N plus r is less than or equal to x less than or equal to minimum of r and s
and is equal to zero otherwise.
The value of N is estimated by the principle of Maximum Likelihood, that is we find the value N cap of N which maximizes p of N. Since N is a discrete random variable, the principle of maxima and minima in calculus cannot be used here. Hence, we proceed as follows.

Lambda of N is equal to p of N divided by p of N minus 1
By substituting and simplifying, we get
N minus r into N minus s whole divided by, N into N minus r minus s plus x
Suppose lambda of N is greater than 1, then
N minus r into N minus s whole divided by, N into N minus r minus s plus x is greater than 1.
On simplification, we get,
N is greater than r into s by x
Therefore p of N is greater than p of N minus if and only if N is greater than r into s by x.

And suppose Lambda is less than 1, then
N minus r into N minus s whole divided by, N into N minus r minus s plus x is less than 1.
On simplification we get, N is less than r into s by x
Therefore p of N is less than p of N minus 1 if and only if N is less than r into s by x.
From above inequalities, we see that p of N reaches the maximum value, which is a function N, when N is approximately equal to r into s by x. hence maximum likelihood estimate of N is given by
N cap is equal to r into s by x.

Now consider the following result:
If X and Y are independent binomial variates with parameters n 1 and p, and n 2 and p, respectively, then show that Probability of X is equal to r given x plus y is equal to n is hypergeometric.

We prove the above result as follows.
We know that x has binomial distribution with parameters n one and p
Hence p of x is equal to n one c x, into p power x into 1 minus p power n one minus x.
Y has binomial distribution with parameters n 2 and p,
P of y is equal to n two c y, p power y into 1 minus p power n two minus y.

Further, since x and y is independent, x plus y is also a binomial variate with parameters n-one plus n-two and p.
Therefore p of x plus y is equal to n one plus n 2 c x plus y into p power x plus y into 1 minus p power n one plus n two minus of x plus y.

Consider the conditional distribution, probability of x is equal to r given x plus y is equal to n
Is equal to probability of x is equal to r intersection x plus y is equal to n divided by probability of x plus y is equal to n
Is equal to probability of x is equal to r intersection y is equal to n minus r, divided by probability of x plus y is equal to n
Since x and y are independent, we can write
Probability of x is equal to r into probability of y is equal to n minus r divided by probability of x plus y is equal to n.

By substituting, we get,
n one c r into p power r into 1 minus p power n one minus r into n two c n minus r into p power n minus r into 1 minus p power n two minus n plus r, divided by n one plus n 2 c n into p power n into 1 minus p power n one plus n 2 minus n.
On simplification, we get
n one c r into n two c n minus r divided by n one plus n 2 c n, which is the probability mass function of hypergeometric distribution with parameters n one, n two and n.
Hence the conditional distribution of x is equal to r given x plus y is equal to n is hypergeometric.

Here's a summary of our learning in this session:

- About hypergeometric distribution
- Mean
- Variance
- Factorial moments
- Approximation to binomial distribution
- Recurrence relation for probabilities
- Application