

5.5: Hypergeometric Distribution

The hypergeometric distribution is a discrete probability distribution used to find the probability of success when there are two outcomes to each trial, and there are a set number of dependent trials. A hypergeometric probability distribution results from a random experiment that meets all of the following requirements. ¹²

1. The procedure has a fixed sample size n taken from a population N .
2. The trials are taken without replacement (dependent).
3. Each trial has exactly two outcomes, the number of successes = a and the number of failures = b . Note $a + b = N$.

If a random experiment satisfies all of the above, the distribution of the random variable X , where X counts the number of successes, is called a hypergeometric distribution, we write $X \sim H(n, a, N)$.

The hypergeometric distribution is $P(X = x) = \frac{{}_a C_x \cdot {}_b C_{n-x}}{{}_N C_n}$, $x = 0, 1, 2, \dots, n$ or a , whichever is smaller. Where n is the sample size taken without replacement from the population size N , x is the number of successes out of n that you are trying to find the probability for, a is the number of a successes and $b = N - a$ is the number of failures out of the population N .

A committee of 5 people is to be formed from 15 volunteers and 6 appointed positions. Compute the probability that the committee will consist of 4 randomly selected volunteers.

Solution

There are a total of $N = 15 + 6 = 21$ people. A sample of $n = 5$ people are selected without replacement. The success is choosing a volunteer, so $a = 15$. Of the 15 volunteers, we are asked to find the probability of selecting $x = 4$. Substitute in each value: $P(X = 4) = \frac{{}_{15}C_4 \cdot {}_6C_1}{{}_{21}C_5} = 0.4025$.

There is a pattern to the numbers in the formula. The top two numbers on the left of the C's add up to the bottom number on the left of C, and the top two numbers on the right of the C's add up to the bottom number on the right of C.

The waiting room for jury selection has 40 people with a college degree and 30 people without. A jury of 12 is selected from this waiting room. What is the probability that exactly 5 of the jury members will have a college degree?

Solution

There are a total of $N = 40 + 30 = 70$ people. A sample of $n = 12$ people are selected without replacement. The success is choosing someone with a college degree, so $a = 40$. Of the 12 jurors we are asked to find the probability of selecting $x = 5$.

Substitute in each value: $P(X = 5) = \frac{{}_{40}C_5 \cdot {}_{30}C_7}{{}_{70}C_{12}} = 0.1259$.

A wallet contains three \$100 bills and five \$1 bills. You randomly choose four bills. What is the probability that you will choose exactly two \$100 bills?

Solution

There are a total of $N = 3 + 5 = 8$ bills. A sample of $n = 4$ bills are selected without replacement. The success is choosing a \$100 bill, so $a = 3$. Of the 3 \$100 bills we are asked to find the probability of selecting $x = 2$. Substitute in each value: $P(X = 2) = \frac{{}_3C_2 \cdot {}_5C_2}{{}_8C_4} = 0.4286$.

Note: the calculator does not have the hypergeometric distribution shortcut. The following online calculator will calculate the probability: <https://homepage.divms.uiowa.edu/~mbognar/applets/hg.html>. The $M = a$ is the number of successes a from our example, or use Excel.

Figure 5-7 may help you decide when to use the "True" for cumulative in Excel.

$P(X = x)$	$P(X \leq x)$	$P(X \geq x)$
Is the same as	Is less than or equal to	Is greater than or equal to
Is equal to	Is at most	Is at least
Is exactly the same as	Is not greater than	Is not less than
Has not changed from	Within	
$=\text{HYPGEOM.DIST}(x, n, a, N, \text{false})$	$= \text{HYPGEOM.DIST}(x, n, a, N, \text{true})$	$= 1 - \text{HYPGEOM.DIST}(x-1, n, a, N, \text{true})$
Where:	$P(X > x)$	$P(X < x)$
x is the value in the question that you are finding the probability for.	More than	Less than
N = Population Size	Greater than	Below
n = Sample Size	Above	Lower than
a = Number of Successes out of N	Higher than	Shorter than
b = Number of Failures out of N	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced
No shortcut for the TI	Smaller	Larger
	$= 1 - \text{HYPGEOM.DIST}(x, n, a, N, \text{true})$	$= \text{HYPGEOM.DIST}(x-1, n, a, N, \text{true})$

Figure 5-7

A bookstore owner examines 5 books from each lot of 25 to check for missing pages. If they find at least 2 books with missing pages, the entire lot is returned. If, indeed, there are 4 books with missing pages, find the probability that the lot will be returned.

Solution

Sometimes the hardest part is determining which of the numbers goes in the right place in the formula. The population size is always the largest number, $N = 25$. A success is what you are trying to find the probability for, which is getting a book with a missing page. The number of successes out of the population is $a = 4$ (the 4 books with the missing pages). This means $b = 25 - 4 = 21$ (the number of books without missing pages). Note $a + b = N$. The sample size is the $n = 5$ is the number of books the owner examines. The owner will send the lot back if X is at least 2 missing pages.

We can stop at $X = 4$ since there are only 4 books with missing pages. Your stopping point will be the smaller of n or a . $P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) = \frac{{}^4C_2 \cdot {}^{21}C_3}{{}^{25}C_5} + \frac{{}^4C_3 \cdot {}^{21}C_2}{{}^{25}C_5} + \frac{{}^4C_4 \cdot {}^{21}C_1}{{}^{25}C_5} = 0.1502 + 0.0158 + 0.0004 = 0.1664$.

In Excel use the formula: $=1 - \text{HYPGEOM.DIST}(1, 5, 4, 25, \text{TRUE})$.

Figure 5-8 shows the online calculator at <https://homepage.divms.uiowa.edu/~mbognar/applets/hg.html>.

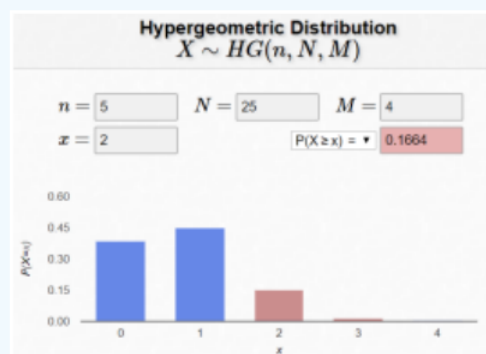


Figure 5-8