Probability and Statistics (IT302) Class No. 21
5th October 2020 Monday 09:45AM - 10:15AM

Hypergeometric Experiment

In general, we are interested in the probability of selecting x successes from the k items labeled successes and n − x failures from the N − k items labeled failures when a random sample of size n is selected from N items. **This is known as a hypergeometric experiment**, that is, one that possesses the following two properties:

1) A random sample of size n is selected without replacement from N items.
2) Of the N items, k may be classified as successes and N − k are classified as failures.

The number X of successes of a hypergeometric experiment is called a hypergeometric random variable. Accordingly, the probability distribution of the hypergeometric variable is called the hypergeometric distribution, and its values are denoted by h(x;N, n, k), since they depend on the number of successes k in the set N from which we select n items.

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Hypergeometric Distribution in Acceptance Sampling

Like the binomial distribution, the hypergeometric distribution finds applications in acceptance sampling, where lots of materials or parts are sampled in order to determine whether or not the entire lot is accepted.

Example 5.8

A particular part that is used as an injection device is sold in lots of 10. The producer deems a lot acceptable if no more than one defective is in the lot. A sampling plan involves random sampling and testing 3 of the parts out of 10. If none of the 3 is defective, the lot is accepted. Comment on the utility of this plan.

Solution : Let us assume that the lot is truly unacceptable (i.e., that 2 out of 10 parts are defective). The probability that the sampling plan finds the lot acceptable is

$$P(X = 0) = \frac{\binom{2}{0}\binom{8}{3}}{\binom{10}{3}} = 0.467$$

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Example 5.8 Contd.

Thus, if the lot is truly unacceptable, with 2 defective parts, this sampling plan will allow acceptance roughly 47% of the time. As a result, this plan should be considered faulty.

Let us now generalize in order to find a formula for $h(x; N, n, k)$. The total number of samples of size n chosen from N items is $\binom{N}{n}$. These samples are assumed to be equally likely. There are $\binom{k}{x}$ ways of selecting x successes from the k that are available, and for each of these ways we can choose the $n - x$ failures in $\binom{N-k}{n-x}$ ways. Thus, the total number of favorable samples among the $\binom{N}{n}$ possible samples is given by $\binom{k}{x}\binom{N-k}{n-x}$. Hence, we have the following definition.

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Hypergeometric Distribution

The probability distribution of the hypergeometric random variable X, the number of successes in a random sample of size n selected from N items of which k are labeled **success** and $N-k$ labeled **failure**, is

$$h(x;N,n,k)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},\qquad \max\{0,n-(N-k)\}\le x\le \min\{n,k\}.$$

The range of x can be determined by the three binomial coefficients in the definition, where x and $n-x$ are no more than k and $N-k$, respectively, and both of them cannot be less than 0. Usually, when both k (the number of successes) and $N-k$ (the number of failures) are larger than the sample size n, the range of a hypergeometric random variable will be $x=0,1,\ldots,n$.

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Example 5.9

Lots of 40 components each are deemed unacceptable if they contain 3 or more defectives. The procedure for sampling a lot is to select 5 components at random and to reject the lot if a defective is found. What is the probability that exactly 1 defective is found in the sample if there are 3 defectives in the entire lot?

Solution : Using the hypergeometric distribution with n = 5, N = 40, k = 3, and x = 1, we find the probability of obtaining 1 defective to be

$$h(1;40,5,3) = \frac{\binom{3}{1}\binom{37}{4}}{\binom{40}{5}} = 0.3011$$

Once again, this plan is not desirable since it detects a bad lot (3 defectives) only about 30% of the time.

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Theorem 5.2

The mean and variance of the hypergeometric distribution $h(x; N, n, k)$ are

$$\mu = \frac{nk}{N} \text{ and } \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}\left(1 - \frac{k}{N}\right).$$

To find the mean of the hypergeometric distribution, we write

$$E(X) = \sum_{x=0}^{n} x \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = k \sum_{x=1}^{n} \frac{(k-1)!}{(x-1)!(k-x)!} \cdot \frac{\binom{N-k}{n-x}}{\binom{N}{n}}$$
$$= k \sum_{x=1}^{n} \frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}}.$$

Since

$$\binom{N-k}{n-1-y} = \binom{(N-1)-(k-1)}{n-1-y} \quad \text{and} \quad \binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N}{n}\binom{N-1}{n-1},$$

letting $y = x - 1$, we obtain

$$E(X) = k \sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{N-k}{n-1-y}}{\binom{N}{n}}$$
$$= \frac{nk}{N} \sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{(N-1)-(k-1)}{n-1-y}}{\binom{N-1}{n-1}} = \frac{nk}{N},$$

since the summation represents the total of all probabilities in a hypergeometric experiment when $N-1$ items are selected at random from $N-1$, of which $k-1$ are labeled success.

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Relationship to the Binomial Distribution

There is an interesting relationship between the hypergeometric and the binomial distribution. As one might expect, if n is small compared to N, the nature of the N items changes very little in each draw. So a binomial distribution can be used to approximate the hypergeometric distribution when n is small compared to N. In fact, as a rule of thumb, the approximation is good when $n/N \leq 0.05$.

Thus, the quantity k/N plays the role of the binomial parameter p. As a result, the binomial distribution may be viewed as a large-population version of the hypergeometric distribution. The mean and variance then come from the formulas

$$\mu = np = \frac{nk}{N} \text{ and } \sigma^2 = npq = n \cdot \frac{k}{N}\left(1 - \frac{k}{N}\right)$$

Comparing these formulas with those of Theorem 5.2, we see that the mean is the same but the variance differs by a correction factor of $(N - n)/(N - 1)$, which is negligible when n is small relative to N.

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Example 5.12

A manufacturer of automobile tires reports that among a shipment of 5000 sent to a local distributor, 1000 are slightly blemished. If one purchases 10 of these tires at random from the distributor, what is the probability that exactly 3 are blemished?

Solution **:** Since $N = 5000$ is large relative to the sample size $n = 10$, we shall approximate the desired probability by using the binomial distribution. The probability of obtaining a blemished tire is 0.2. Therefore, the probability of obtaining exactly 3 blemished tires is $h(3; 5000, 10, 1000) \approx b(3; 10, 0.2) = 0.8791 - 0.6778 = 0.2013$.

On the other hand, the exact probability is $h(3; 5000, 10, 1000) = 0.2015$.

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

The hypergeometric distribution can be extended to treat the case where the N items can be partitioned into k cells A1,A2, . . . , Ak with a1 elements in the first cell, a2 elements in the second cell, . . . , ak elements in the kth cell. We are now interested in the probability that a random sample of size n yields x1 elements from A1, x2 elements from A2, . . . , and xk elements from Ak. Let us represent this probability by

$$f(x_1, x_2, \ldots, x_k; a_1, a_2, \ldots, a_k, N, n).$$

To obtain a general formula, we note that the total number of samples of size n that can be chosen from N items is still $\binom{N}{n}$. There are $\binom{a_1}{x_1}$ ways of selecting x_1 items from the items in A_1, and for each of these we can choose x_2 items from the items in A_2 in $\binom{a_2}{x_2}$ ways. Therefore, we can select x_1 items from A_1 and x_2 items from A_2 in $\binom{a_1}{x_1}\binom{a_2}{x_2}$ ways. Continuing in this way, we can select all n items consisting of x_1 from A_1, x_2 from A_2, . . . , and x_k from A_k in

$$\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k} \quad \text{ways.}$$

The required probability distribution is now defined as follows.

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Multivariate Hypergeometric Distribution

If N items can be partitioned into the k cells A_1, A_2, . . . , A_k with a_1, a_2, . . . , a_k elements, respectively, then the probability distribution of the random variables X_1, X_2, . . . , X_k, representing the number of elements selected from A_1, A_2, . . . , A_k in a random sample of size n, is

$$f(x_1, x_2, \ldots, x_k; a_1, a_2, \ldots, a_k, N, n) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k}}{\binom{N}{n}}$$

$$\text{with } \sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} a_i = N.$$

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Example 5.13

A group of 10 individuals is used for a biological case study. The group contains 3 people with blood type O, 4 with blood type A, and 3 with blood type B. What is the probability that a random sample of 5 will contain 1 person with blood type O, 2 people with blood type A, and 2 people with blood type B?

Solution : Using the extension of the hypergeometric distribution with $x_1 = 1$, $x_2 = 2$, $x_3 = 2$, $a_1 = 3$, $a_2 = 4$, $a_3 = 3$, N = 10, and n = 5, we find that the desired probability is

$$f(1, 2, 2; 3, 4, 3, 10, 5) = \frac{\binom{3}{1}\binom{4}{2}\binom{3}{2}}{\binom{10}{5}} = \frac{3}{14}$$

Source : Probability & Statistics for Engineers & Scientists, by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, 9th Edition, Prentice Hall.

Additional Material

Assumptions of the Hypergeometric Distribution

- It is a discrete distribution.
- Sampling is done *without replacement.*
- The number of objects in the population, N, is finite and known.
- Each trial has exactly two possible outcomes: success and failure.
- Trials are not independent
- X is the number of successes in the n trials

Source: http://www2.aueb.gr/users/koundouri/resees/uploads/Chapter05.ppt

Hypergeometric Distribution

- Probability function
  - N is population size
  - n is sample size
  - A is number of successes in population
  - x is number of successes in sample

$$P(x) = \frac{\left({}_{A}C_{x}\right)\left({}_{N-A}C_{n-x}\right)}{{}_{N}C_{n}}$$

Mean value

$$\mu = \frac{A \cdot n}{N}$$

- Variance and Standard Deviation

$$\sigma^2 = \frac{A(N-A)n(N-n)}{N^2(N-1)}$$

$$\sigma = \sqrt{\sigma^2}$$

Source: http://www2.aueb.gr/users/koundouri/resees/uploads/Chapter05.ppt

Hypergeometric Distribution: Probability Computations

N = 24
X = 8
n = 5

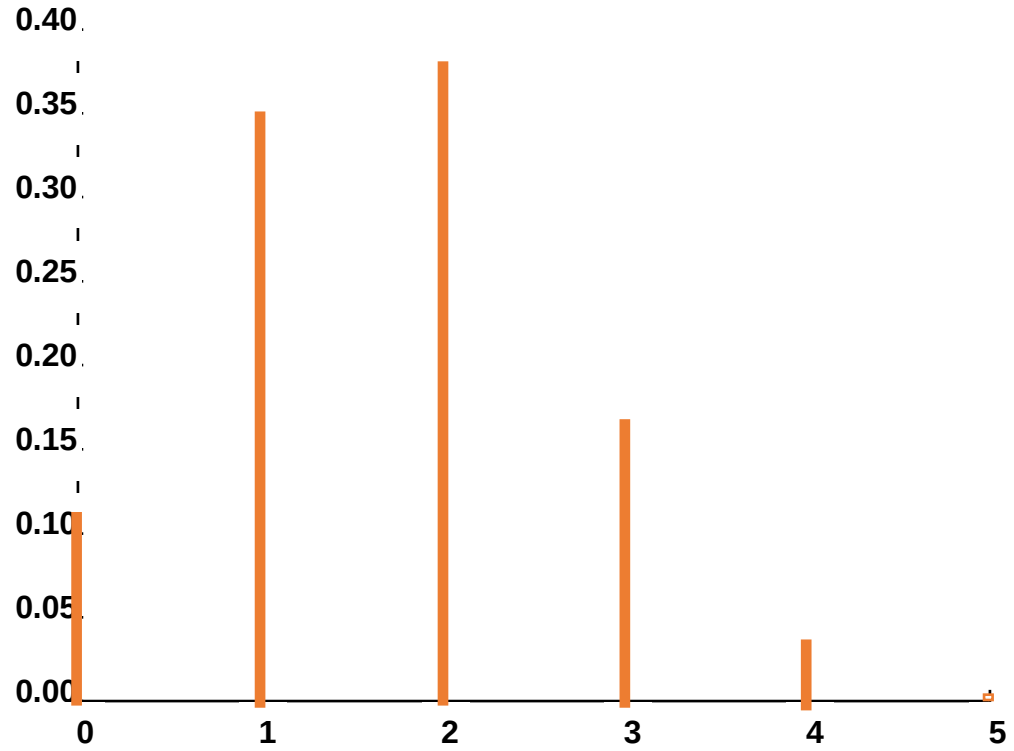
x	P(x)
0	0.1028
1	0.3426
2	0.3689
3	0.1581
4	0.0264
5	0.0013

$$P(x=3) = \frac{\left({}_{A}C_{x}\right)\left({}_{N-A}C_{n-x}\right)}{{}_{N}C_{n}}$$

$$= \frac{\left({}_{8}C_{3}\right)\left({}_{24-8}C_{5-3}\right)}{{}_{24}C_{5}}$$

$$= \frac{(56)(120)}{42{,}504}$$

$$= .1581$$



Hypergeometric Distribution: Graph

Source: http://www2.aueb.gr/users/koundouri/resees/uploads/Chapter05.ppt

The Hypergeometric Distribution and the Binomial Distribution

- Because the hypergeometric distribution is described by three parameters N, A and n, it is practically impossible to create tables for easy use.
- The binomial (which has tables) is an acceptable approximation, if $n < 5\%\ N$. Otherwise it is not.

Hypergeometrıc Dıstrıbutıon

$$f(X / A, B, n) = \frac{\binom{A}{x} . \binom{B}{n-x}}{\binom{A+B}{n}}$$

n= sample size

A+B=population size

A=successes in population

X=number of successes in sample

Mean , Variance and Standard Deviation

$$\mu = E(X) = \frac{n \cdot A}{A + B}$$

$$Var(X) = \frac{n \cdot A \cdot B}{(A+B)^2} \cdot \frac{A+B-n}{A+B-1}$$

Source: http://kisi.deu.edu.tr/s.gumusoglu/dosyalar/hypergeometric%20distribution.ppt

Hypergeometrıc Dıstrıbutıon Contd.

APPROXİMATİONS

Binomial Approximation Requariments :

If A+B=N and n ≤0,05N , Binomial can be used instead of hypergeometric distribution

Poisson Approximation Requariments:

If $n \leq 0{,}05N$ $n \geq 20$ $P \leq 0{,}05$

Poisson can be used instead of hypergeometric distribution

Source: http://kisi.deu.edu.tr/s.gumusoglu/dosyalar/hypergeometric%20distribution.ppt

Hypergeometrıc Dıstrıbutıon

Example 1 :A carton contains 24 light bulbs, three of which are defective. What is the probability that, if a sample of six is chosen at random from the carton of bulbs, x will be defective?

$$P(X = x) = \frac{\binom{3}{x}.\binom{21}{6-x}}{\binom{24}{6}} \qquad P(X = 0) = \frac{\binom{3}{0}.\binom{21}{6}}{\binom{24}{6}} = 0{,}40316$$

That is no defective

$$P(X = 3) = \frac{\binom{3}{3}.\binom{21}{3}}{\binom{24}{6}} = 0{,}00988$$

That is 3 will be defective.

Source: http://kisi.deu.edu.tr/s.gumusoglu/dosyalar/hypergeometric%20distribution.ppt

Example 2: Suppose that 7 balls are selected at random without replacement from a box containing 5 red balls and 10 blue balls .If X denotes the proportion of red balls in the sample, what are the mean and the variance of X ?

A=5 red B=10 blue A+B =15 n=7

$$Var(X) = \frac{n \cdot A \cdot B}{(A+B)^2} \cdot \frac{A+B-n}{A+B-1}$$

$$= \frac{7 \cdot 5 \cdot 10}{15^2} \cdot \frac{15-7}{15-1} = 0{,}8888$$

$$E(X) = \frac{n \cdot A}{A+B} = \frac{7 \cdot 5}{15} = 2{,}33$$

Source: http://kisi.deu.edu.tr/s.gumusoglu/dosyalar/hypergeometric%20distribution.ppt

Example4:Suppose that a shipment contains 5 defective items and 10 non defective items .If 7 items are selected at random without replacement , what is the probability that at least 3 defective items will be obtained?

N=15 (5 defective , 10 nondefective) n=7

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - [P(0) + P(1) + P(2)] = 0{,}4267$$

$$P(0) = \frac{\binom{5}{0} \cdot \binom{10}{7}}{\binom{15}{7}} = 0{,}0186$$

$$P(1) = \frac{\binom{5}{1} \cdot \binom{10}{6}}{\binom{15}{7}} = 0{,}1631$$

$$P(2) = \frac{\binom{5}{2} \cdot \binom{10}{5}}{\binom{15}{7}} = 0{,}3916$$

Source: http://kisi.deu.edu.tr/s.gumusoglu/dosyalar/hypergeometric%20distribution.ppt

Example 3 : If a random variable X has a hyper geometric distribution with parameters A=8 , B=20 and n, for what value of n will Var(x) be maximum ?

$$Var(X) = \frac{n \cdot A \cdot B}{(A+B)^2} \cdot \frac{A+B-n}{A+B-1} = \frac{n \cdot 8 \cdot 20}{(8+20)^2} \cdot \frac{8+20-n}{8+20-1} =$$

$$= \frac{160n}{28^2} \cdot \frac{(28-n)}{27} = 0$$

n=28 or n=0 for variance to be maximum

Source: http://kisi.deu.edu.tr/s.gumusoglu/dosyalar/hypergeometric%20distribution.ppt