Lesson 2.6

The Hypergeometric Distribution

Learning Outcomes

At the end of the lesson, students must be able to

1. describe the hypergeometric distribution,
2. derive the probability mass functions of a random variable with a hypergeometric distribution,
3. compute probabilities associated with a random variable with a hypergeometric distribution, and
4. compute the mean and variance of a random variable with a hypergeometric distribution.

Introduction

Suppose 6 statisticians and 19 mathematicians are attending a conference and all 25 names are put in a hat and 5 names are randomly picked without replacement. What is the probability that 4 statisticians and 1 mathematician are picked?

One of the keys here is that we are picking these names without replacement and by without replacement we mean once a name has been picked once it is removed and cannot be picked again. So that implies that these trials are not independent. Thus, the binomial distribution does not apply here.

Before going any further, let us compute the probability that 4 statisticians and 1 mathematician are picked. Assuming any sample of 5 names is equally likely, then

$$\begin{aligned} P(\text{4 stat and 1 math}) &= \frac{\text{number of samples with 4 stat and 1 math}}{\text{total number of samples of size 5}} \\ &= \frac{\binom{6}{4}\binom{19}{1}}{\binom{25}{5}} \\ &\approx 0.0054 \end{aligned}$$

Since the sampling is dome without replacement, the probability of success changes from trial to trial. For example, the probability that the first draw resulted to a statistician is $\frac{6}{25}$. Now, suppose the first draw was a statistician, then the probability that another statistician is picked in the second draw is $\frac{5}{24}$. Hence, the probability of selecting a statistician is not constant from one draw to another. Hence, the binomial distribution does not apply in this type of problem.

Definition:

Suppose we are randomly sampling n objects without replacement from a source or container that contains k successes and $N-k$ failures. Let Y be the random variable representing the number of successes in the sample. Then Y has a hypergeometric distribution with probability mass function given by

$$P(Y=y) = \frac{\binom{k}{y}\binom{N-k}{n-y}}{\binom{N}{n}}$$

and we write $Y \sim Hyper(N, n, k)$.

Example 1:

An urn contains ten marbles, of which five are green, two are blue, and three are red. Three marbles are to be drawn from the urn, one at a time without replacement. What is the probability that all three marbles drawn will be green?

Solution:

Let Y denote the number of green marbles. We have $N = 10, k = 5, n = 3$. Thus, $Y \sim Hyper(10, 3, 5)$ and

$$P(Y=3) = \frac{\binom{5}{3}\binom{5}{0}}{\binom{10}{3}} \approx 0.083$$

This probability can also be obtained using the R command *dhyper(3, 5, 5, 3)*.

Example 2:

A supplier ships parts to a company in lots of 1000 parts. Suppose a lot contains 100 defective parts and 900 non-defective parts. An operator selects 10 parts at random and without replacement. What is the probability he selects no more than 2 defective parts?

Solution:

Let X denote the number of defective parts. We have $N = 1000, k = 100, n = 10$. Thus, $Y \sim Hyper(1000, 10, 100)$ and

$$\begin{aligned} P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= \frac{\binom{100}{0}\binom{900}{10}}{\binom{1000}{10}} + \frac{\binom{100}{1}\binom{900}{9}}{\binom{1000}{10}} + \frac{\binom{100}{2}\binom{900}{8}}{\binom{1000}{10}} \\ &\approx 0.9308 \end{aligned}$$

This probability can be obtained using the R command *phyper(2, 100, 900, 10)*.

Theorem: If $Y \sim Hyper(N, n, k)$, then

$$\begin{aligned} E(Y) &= n\left(\frac{k}{N}\right) \\ V(Y) &= n\left(\frac{k}{N}\right)\left(\frac{N-k}{N}\right)\left(\frac{N-n}{N-1}\right) \end{aligned}$$

Proof: Left as an exercise.

Remark:

The motivation for the hypergeometric distribution should remind us of the underlying framework for the binomial; i.e., we record the number of "successes" out of n trials. The difference here is that (a) the population size N is finite and (b) sampling is done without replacement. To understand this let

$$p = \frac{k}{N} = \text{proportion of "successes" in the population}$$

Because sampling from the population is done without replacement, the value of p changes from trial to trial. This violates the Bernoulli trial assumptions, so technically the binomial model does not apply. However, one can show mathematically that as $\frac{k}{N} \to p$

$$\lim_{N\to\infty} \frac{\binom{k}{y}\binom{N-k}{n-y}}{\binom{N}{n}} = \binom{n}{y} p^y (1-p)^{n-y}$$

This result implies that if the population size N is "large," the $hyper(N, n, k)$ distribution and the $binom(n, p = \frac{k}{N})$ distribution should be very close to each other even when one samples without replacement. Of course, if one samples from a population with replacement, then $p = \frac{k}{N}$ remains fixed and hence the binomial model applies regardless of how large N is.

The above discussion provides us to approximate the hypergeometric distribution using a binomial distribution. If sampling is with replacement, then Y possesses an approximate binomial distribution when N is large and n is relatively small, preferably $\frac{n}{N} \leq 0.10$.

Example 3:

Batteries of a certain gadget are shipped in boxes of 500. The lab technician takes a random sample of 4 batteries from a box upon receipt. If 2 or more batteries are nonworking, the box is rejected and returned for replacement. If 10% of the batteries is usually nonworking, find the probability of rejecting a box.

Solution:

Let Y be the number of nonworking batteries.

a) Using the hypergeometric distribution

If 10% of the batteries is usually nonworking , then $k = 0.1 \times 500 = 50$ hence, $Y \sim hyper(500, 4, 50)$.

$$\begin{aligned} P(\text{reject the box}) = P(Y \geq 2) &= 1 - P(Y < 2) \\ &= 1 - [P(Y = 0) + P(Y = 1)] \\ &= 1 - \left[\frac{\binom{50}{0}\binom{450}{4}}{\binom{500}{4}} + \frac{\binom{50}{1}\binom{450}{3}}{\binom{500}{4}} \right] \\ &\approx 0.0516 \end{aligned}$$

b) Using the binomial approximation

we have $\frac{n}{N} = \frac{4}{500} = 0.008 < 0.10$ so a binomial approximation is warranted with $p = \frac{50}{500} = 0.10$. Thus, $Y \sim binom(4, 0.10)$.

$$\begin{aligned} P(\text{reject the box}) = P(Y \geq 2) &= 1 - P(Y < 2) \\ &= 1 - [P(Y = 0) + P(Y = 1)] \\ &= 1 - \left[\binom{4}{0}(0.10)^0(1 - 0.10)^4 + \binom{4}{1}(0.10)^1(1 - 0.10)^3 \right] \\ &\approx 0.0523 \end{aligned}$$

Example 4:

Seeds are often treated with fungicides to protect them in poor draining, wet environments. A small-scale trial, involving five treated and five untreated seeds, was conducted prior to a large-scale experiment to explore how much fungicide to apply. The seeds were planted in wet soil, and the number of emerging plants were counted. If the solution was not effective and four plants actually sprouted, what is the probability that

a) all four plants emerged from treated seeds?
b) three or fewer emerged from treated seeds?
c) at least one emerged from untreated seeds?

Solution: Left as a classroom exercise.