3.3.5 Hypergeometric Random Variables

Our next type of discrete random variable is the hypergeometric random variable.

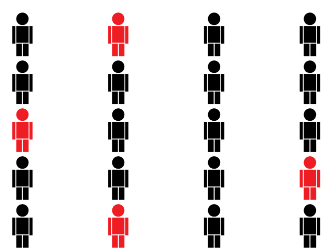
Example 8: Consider the social deduction game werewolf. The game involves a group of villagers that have werewolves hiding among them. The game takes place during a series of days and nights. During the night all villagers keep their eyes closed until specific roles are called to "wake up" by the narrator of the game.

When the werewolves wake up during the night may eliminate one person. After the night is over the narrator says "everyone wake up" and the day begins. During the day the villagers may eliminate a person in hopes of taking out a werewolf.

The game ends when all werewolves are eliminated (villagers win) or three people remain in the village and one is a werewolf (werewolves win). There are also several other roles such as doctor, witch, tanner, but we will neglect these for this problem.

Let's now consider you are playing werewolf with 20 friends. 4 of your friends are designated as werewolves. You are given the most prestigious role of all; the villages "super statistician". During the first night you have a chance to sample four villagers. All the players you select are automatically eliminated.

(a) Let X be the number of werewolves in your super statistician sample. What is the probability that $X = 1$?



Here we define our sample space as the total number of ways we can sample four people from your 20 friends.

$$\begin{aligned} \#\text{ of ways to sample 4 friends from 20} &= \binom{20}{4} \\ &= 4845 \end{aligned}$$

Next we want the total number of samples that would result in you choosing 1 werewolf. This means we are interested in the number of ways you pull 3 villagers **and** 1 werewolf. We will use the multiplication rule for counting here.

$$\begin{aligned} \#\text{ of ways to pull 3 villagers } \textbf{AND} \text{ 1 werewolf} &= \binom{16}{3}\binom{4}{1} \\ &= 2240 \end{aligned}$$

Now we can determine the probability of having exactly one werewolf in our sample.

$$\begin{aligned} P(X = 1) &= \frac{\binom{16}{3}\binom{4}{1}}{\binom{20}{4}} \\ &= \frac{2240}{4845} \\ &= 0.4623323 \end{aligned}$$

(b) Whats the probability of having one or more werewolves in your super statistician sample?

We will use complimentary probability here

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - \frac{\binom{16}{4}\binom{4}{0}}{\binom{20}{4}} \\ &= 0.624355005 \end{aligned}$$

(c) Run five simulations to estimate the expected number of werewolves you would have in your sample.

1. Here we are interested in simulating the expected number werewolves we will get in the super statistician sample.
2. We are sampling without replacement, so we must ignore repeated numbers.
3. 4 out of the 20 friends are werewolves. We will assign the numbers 0-19 to werewolves, and 20-99 to villagers.
4. Not we will run several simulations.

Simulation Number	Simulation	X
1	52 93 63 00	1
2	64 73 10 34	1
3	73 53 45 95	0
4	58 94 23 13	2
5	52 79 50 22	1

5. We would estimate the expected number of werewolves in our sample to be the average of X over the five simulations. This would be $E(X) \approx 1$ werewolf. Running 100000 simulations in R gives $E(X) \approx 0.8001737$.

   The theoretical expected value for a hypergeometric random variable is $E(X) = n\frac{r}{N}$, where n is the sample size, N is the population size, and r is the number of successes in the population. For this example we have $E(X) = 4\left(\frac{4}{20}\right) = \frac{16}{20} = 0.8$ werewolves. Proving the theoretical expectation is left as an exercise.

(d) Create a probability distribution table for X

x	0	1	2	3	4
$P(X = x)$	0.375644995	0.462332301	0.148606811	0.013209494	0.000206398

- $P(X = 0) = \frac{\binom{16}{4}\binom{4}{0}}{\binom{20}{4}}$
- $P(X = 1) = \frac{\binom{16}{3}\binom{4}{1}}{\binom{20}{4}}$
- $P(X = 2) = \frac{\binom{16}{2}\binom{4}{2}}{\binom{20}{4}}$
- $P(X = 3) = \frac{\binom{16}{1}\binom{4}{4}}{\binom{20}{4}}$
- $P(X = 4) = \frac{\binom{16}{0}\binom{4}{4}}{\binom{20}{4}}$

(e) Calculate $Var(X)$ using the probability distribution table for X.

We know that $E(X) = 0.8$ (Note we could have also calculated this using the probability distribution table). Now lets calculate $E(X^2)$.

$$\begin{aligned} E(X^2) &= \sum_{x=0}^{4} p(x) \cdot x^2 \\ &= p(0) \cdot 0 + p(1) \cdot 1^2 + p(2) \cdot 2^2 + p(3) \cdot 3^3 + p(4) \cdot 4^2 \\ &= 1.178947 \end{aligned}$$

Then we can calculate $Var(X)$

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 \\ &= 1.178947 - (0.8)^2 \\ &= 0.5389474 \end{aligned}$$

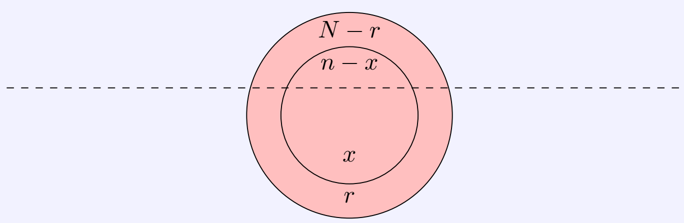
However, like with every random variable so far we have a shortcut. There is a general formula for the theoretical variance of a hypergeometric random variable.

$$Var(X) = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

This is for sampling n elements from a population of N elements, where the population has r successes. For this particular example $n = 4$, $N = 20$, and $r = 4$.

$$\begin{aligned} Var(X) &= 4\left(\frac{4}{20}\right)\left(1 - \frac{4}{20}\right)\left(\frac{20-4}{20-1}\right) \\ &= 0.5389474 \end{aligned}$$

Hypergeometric Random Variables Let X be a hypergeometric variable with parameters N, n, and r. X describes taking a sample of n from the population of size N that contains r 'successes'. We can draw this scenario nicely using an Euler diagram.



1. **Probability Mass Function:** $P(X = x) = \dfrac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, 2, \ldots, n$

2. **Expected Value:** $E(X) = n\left(\dfrac{r}{p}\right)$

3. **Variance:** $Var(X) = n\left(\dfrac{r}{N}\right)\left(1 - \dfrac{r}{N}\right)\left(\dfrac{N-n}{N-1}\right)$

At first glance the formulas for $E(X)$ and $Var(X)$ for hypergeometric random variables look a little intimidating. However, they draw a striking resemblance to that of a binomial random variable.

You can think of $\left(\frac{r}{N}\right) = p^{\star}$. This is the proportion of 'successes' in the population. Then

$$\begin{aligned} E(X) &= np^{\star} \\ Var(X) &= np^{\star}(1-p^{\star})\left(\frac{N-n}{N-1}\right) \end{aligned}$$

The expectation formula is identical, and the variance formula is similar but with $\left(\frac{N-n}{N-1}\right)$ on the end. $\left(\frac{N-n}{N-1}\right)$ is called the *finite population correction factor*. We will talk about this more later. Notice that for $n << N$, $\left(\frac{N-n}{N-1}\right) \approx 1$. It turns out for $n << N$ We can approximate hypergeometric random variables using binomial random variables.

Example 9: Lets revisit the *Unstable Unicorns* example from chapter two. In a game of *Unstable Unicorns* each player is dealt 7 random cards, from a deck that contains 20 instant cards, 25 upgrade cards, 25 downgrade cards, 5 magic cards, 10 magical unicorn cards, and 15 basic unicorn cards. You are dealt 7 cards from the top of the deck. Let X be the number of unicorn cards you draw.

(a) What is $P(X = 2)$?

X is a hypergeometric random variable with $n = 7$, $N = 20 + 25 + 25 + 5 + 10 + 15 = 100$, and $r = 25$.

$$\begin{aligned} P(X=2) &= P(X=x) = \frac{\binom{25}{2}\binom{75}{5}}{\binom{100}{7}} \\ &= 0.323460711 \end{aligned}$$

(b) What is $P(X < 7)$? We will use complimentary probability here

$$\begin{aligned} P(X<7) &= 1 - P(X=7) \\ &= 1 - \frac{\binom{25}{7}\binom{75}{0}}{\binom{100}{7}} \\ &= 0.99996997 \end{aligned}$$

It's pretty likely you won't get a hand full of unicorns... assuming no one is cheating.

(c) How many unicorn cards would you expect to have in your hand?

$$\begin{aligned} E(X) &= n\left(\frac{r}{N}\right) \\ &= 7\left(\frac{25}{100}\right) = 1.75 \end{aligned}$$

You would expect 1.75 unicorns in your hand.

(d) Calculate $Var(X)$

$$\begin{aligned} Var(X) &= n\left(\frac{r}{N}\right)\left(1-\frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \\ &= 7\left(\frac{25}{100}\right)\left(1-\frac{25}{100}\right)\left(\frac{100-7}{100-1}\right) = 1.232955 \end{aligned}$$

$Var(X) = 1.232955$ unicorns^2. The standard deviation is $SD(X) = \sqrt{1.232955} = 1.110385$ unicorns.

