§ 5.4 Hypergeometric distribution

Example Consider the experiment where we draw n balls from an urn which contains N balls in total, r of which are red. Let X count the number of red balls we get when

① Sampling with replacement

② Sampling without replacement.

Find the prob. mass function of X in each case.

① $X \sim \text{Bin}\left(n, \frac{r}{N}\right)$, so $P(X=k) = \binom{n}{k} \left(\frac{r}{N}\right)^{k} \left(1-\frac{r}{n}\right)^{n-k}$

for $k = 0, 1, \ldots, n$.

② $P(X=k) = \dfrac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$ for $0 \le k \le r$ and $0 \le n-k \le N-r$

Def A random variable X has the hypergeometric distribution with parameters r, N, n if its prob. mass function is given by

$$P(X=k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$$

where k is restricted by $0 \le k \le r$ and $0 \le n-k \le N-r$.

Shorthand: $X \sim \text{HyperGeo}(r, N, n)$

R commands: $P(X=k)$: dhyper(k, r, N-r, n)

$P(X \le k)$: phyper(k, r, N-r, n)

Remark $X \sim \text{HyperGeo}(n, N, r)$ models the number of elements of type S in a sample of size n from a population of size N which contains r elements of type S and $N-r$ elements of type F, where the sampling is done without replacement.

Example Let $X \sim \text{HyperGeo}(r, N, n)$. Show $E[X] = \frac{rn}{N}$.

We think of X as counting the number of red balls in a sample of size n (without replacement) from an urn which contains r red balls and $N-r$ non-red balls.

Let $I_1, \dots, I_n$ be random variables where

$$I_k = \begin{cases} 1 & \text{if draw } k \text{ is red} \\ 0 & \text{if draw } k \text{ is not red.} \end{cases}$$

Then $X = I_1 + \cdots + I_n$ and s

$$E[X] = E[I_1] + \cdots + E[I_n].$$

Note $E[I_1] = P(I_1 = 1) = \frac{r}{N}$. Further,

$$\begin{aligned} E[I_2] &= P(I_2 = 1) \\ &= P(I_2 = 1 \mid I_1 = 1)\, P(I_1 = 1) + P(I_2 = 1 \mid I_1 = 0)\, P(I_1 = 0) \\ &= \frac{r-1}{N-1} \cdot \frac{r}{N} + \frac{r}{N-1} \cdot \frac{N-r}{N} \\ &= \frac{r(r-1) + r(N-r)}{N(N-1)} \\ &= \frac{r(r-1+N-r)}{N(N-1)} \\ &= \frac{r}{N} \end{aligned}$$

In fact, $E[I_1] = E[I_2] = \cdots = E[I_n] = \frac{r}{N}$. So $E[X] = \frac{rn}{N}$.

Problem 1. Suppose in a population of 1000 deer at a national park, 200 are captured by the wildlife service. These deer are tagged, say with some identifying mark, and are released back into the park with the rest of the population. After a period of time, a new sample of 50 deer are re-captured. Let X count the number of tagged deer in this group of 50.

a. Does X have the hypergeometric distribution? If so, what are the parameters?

b. Find the probability that there are exactly 10 tagged deer in our re-capture sample.

c. Find the probability that there are at least 10 tagged deer in our re-capture sample.

d. Find the probability that there are between 5 and 15 tagged deer in our re-capture sample.

e. Find the expected number of tagged deer in our re-capture sample.

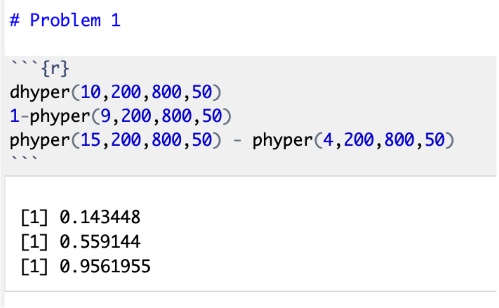
(a) Yes, $X \sim \text{HyperGeo}(200, 1000, 50)$

(b) $P(X=10) = \frac{\binom{200}{10}\binom{800}{40}}{\binom{1000}{50}} \approx 0.143448$

(c) $P(X \geq 10) = 1 - P(X \leq 9) \approx 0.559144$

(d) $P(5 \leq X \leq 15) = P(X \leq 15) - P(X \leq 4) \approx 0.9561955$

(e) $E[X] = \frac{(200)(50)}{1000} = 10$



~~~
# Problem 1

```{r}
dhyper(10,200,800,50)
1-phyper(9,200,800,50)
phyper(15,200,800,50) - phyper(4,200,800,50)
```

[1] 0.143448
[1] 0.559144
[1] 0.9561955
~~~

Problem 2. Suppose that a batch of 100 items contains 6 that are defective and 94 that are not defective. If X is the number of defective items in a randomly drawn sample of 10 items from the batch, find

a. $P(X = 0)$

b. $P(X > 2)$

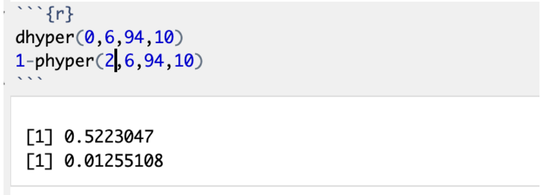
c. $E[X]$

$X \sim \text{HyperGeo}(6, 100, 10)$

(a) $P(X=0) = \frac{\binom{6}{0}\binom{94}{10}}{\binom{100}{10}} \approx 0.5223047$

(b) $P(X > 2) = 1 - P(X \leq 2) \approx 0.01255108$

(c) $E[X] = \frac{(6)(10)}{100} = 0.6$



~~~
```{r}
dhyper(0,6,94,10)
1-phyper(2,6,94,10)
```

[1] 0.5223047
[1] 0.01255108
~~~

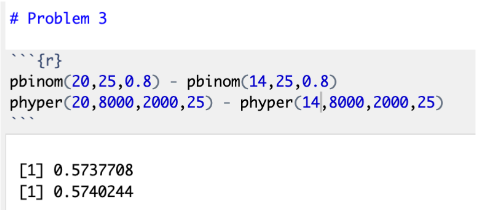
Problem 3. In a population of 10,000 people, 8000 are in favor of a proposal for a new law. Suppose that a poll of 25 people is taken. Find the probability that between 15 and 20 people in the poll are in favor of the proposal, assuming (a) sampling is done with replacement and (b) sampling is done without replacement. What do you notice? Make a conjecture about when the binomial and hypergeometric distributions give similar probabilities.

(a) $X \sim \text{Bin}\left(25, \frac{8000}{1000}\right)$

$P(15 \leq X \leq 20) \approx 0.5737708$

(b) $Y \sim \text{HyperGeo}(8000, 10000, 25)$

$P(15 \leq X \leq 20) \approx 0.5740244$



~~~
# Problem 3

```{r}
pbinom(20,25,0.8) - pbinom(14,25,0.8)
phyper(20,8000,2000,25) - phyper(14,8000,2000,25)
```

[1] 0.5737708
[1] 0.5740244
~~~