

## Hypergeometric Distribution

A random variable  $Y$  has a hypergeometric distribution if

- A sample of size  $n$  is selected without replacement from a population of size  $N$ .
- Each member of the population belongs to one of two groups; success or failure. The number of successes in the population is denoted as  $r$  and therefore the number of failures in the population is  $N - r$ .
- The random variable  $Y$  is the total number of successes in the sample of size  $n$ .
- The parameters for the hypergeometric random variable  $Y$  are the sample size  $n$ , the population size  $N$  and the number of successes in the population  $r$ .
- The probability distribution function of  $Y$  is

$$P(Y = y) = p(y) = \frac{\binom{r}{y} \binom{N-r}{n-y}}{\binom{N}{n}} \quad y = 0, 1, \dots, n \text{ where } y \leq r \text{ and } n - y \leq N - r$$

- The theoretical mean of the hypergeometric random variable  $Y$  is

$$\mu = E(Y) = \frac{nr}{N}$$

- The theoretical variance of the hypergeometric random variable  $Y$  is

$$\sigma^2 = V(Y) = n \left( \frac{r}{N} \right) \left( \frac{N-r}{N} \right) \left( \frac{N-n}{N-1} \right)$$

Working with hypergeometric random variables in R.

To find a probability  $P(Y = y) = p(y)$  for a single value  $y$ , the command in R is

```
dhyper(y,r,N-r,n)
```

To find the probability  $P(Y \leq y)$ , use the sum command to add up all  $p(y)$  values for  $y$  between and including 0 and  $y$ .

```
sum(dhyper(0:y,r,N-r,n))
```

To find the probability  $P(y_1 \leq Y \leq y_2)$ , use the sum command to add up all  $p(y)$  values for  $y$  between and including  $y_1$  and  $y_2$ .

```
sum(dhyper(y1:y2,r,N-r,n))
```

To find the probability  $P(Y \geq y)$ , use the sum command to add up all  $p(y)$  values between and including  $y$  and  $n$ .

```
sum(dhyper(y:n,r,N-r,n))
```

Problems.

1. How is the probability distribution function  $p(y)$  derived?
2. Let the quantity  $r/N = p$ . Derive the expected value and variance of a hypergeometric distribution in terms of  $p$ ,  $n$ , and  $N$ .
3. A box contains 40 balls: 10 red, 15 yellow, and 15 green. A sample of 3 balls is taken from this box without replacement.
  - (a) What is the probability that all three balls will be yellow?
  - (b) What is the probability that exactly two out of three balls will be yellow?
  - (c) What is the expected number of yellow balls in the sample?
  - (d) What is the variance of the number of yellow balls in the sample?
4. Crates of eggs are inspected for blood clots. A sample of three eggs are selected without replacement from a crate of 120 eggs.
  - (a) What is the probability that exactly one out of the three eggs will have a blood clot if the crate contains a total of 10 eggs with blood clots?
  - (b) Use R to calculate the probability that exactly one of the three eggs will have a blood clot if the crate contains a total of  $r$  eggs with blood clots for all possible values of  $r$ .
  - (c) For what value of  $r$  is the probability in part (b) maximized?
5. On any MP3 player, the number of songs from any particular artist that appear in the first  $n$  songs of a  $N$  song playlist has a hypergeometric distribution where  $r$  is the total number of songs from that artist in the playlist. In my favorites playlist of 240 songs, I have 14 songs by the artist Queen.
  - (a) If I listen to the first  $n = 60$  songs on my favorites playlist, find the probability that I will hear six songs by Queen.
  - (b) Use R to determine the probability distribution function for the number of songs from Queen that will appear in the first  $n = 60$  songs on my favorites playlist.
  - (c) In one particular shuffle, I heard all 14 Queen songs in the first 60 songs of the shuffle. Should I question the randomness of the shuffle feature?
  - (d) If each shuffle is independent, in how many shuffles out of 100 total (listening to the first  $n = 60$  songs) should I expect to hear six songs by Queen?
  - (e) If each shuffle is independent, how many shuffles would I need to perform in order to expect to hear all 14 songs by Queen (listening to the first  $n = 60$  songs) just once?