Hypergeometric Distribution

Definition: Binomial Distribution

To say that a random variable X has a *Hypergeometric* distribution with parameters N, r, and n, denoted:

$$X \sim \text{HyperGeom}(n, p)$$

means that:

1. The underlying experiment is composed of n repeated Bernoulli trials of selecting elements from a population of finite size N.
2. The n trials are independent.
3. There are r elements in the population that result in success when selected.
4. Every subset of n elements from the population has an equal probability of being selected.
5. X counts the number of successes resulting from the n trials.

Note that a Hypergeometric distribution for selection from a population implies *no* replacement.

Example: Hypergeometric Distributions

1. Select (without replacement) 10 balls from an urn that has 30 red balls and 20 blue balls: $X =$ the number of selected red balls.

   $$X \sim \text{HyperGeom}(50, 30, 10)$$

2. Poll n different voters at random from the whole pool of N registered voters, r of which support a certain presidential candidate: $Y =$ the number of polled voters that support the candidate.

   $$X \sim \text{HyperGeom}(N, r, n)$$

Theorem

Let X be a random variable with a Hypergeometric distribution with parameters N, r, and n such that $x \leq r$ and $n - x \leq N - r$, and let $p = \frac{r}{n}$:

- $f_X(x) = \begin{cases} \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} & x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$
- $E(X) = \frac{nr}{N} = np$
- $V(X) = np(1-p)\left(\frac{N-n}{N-1}\right)$

Proof. For $P(X = x)$, select any x of r: $\binom{r}{x}$, then select any $n - x$ of $N - r$: $\binom{N-r}{n-x}$. The total number of possible selections is $\binom{N}{n}$. Therefore:

$$f_X(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

■

Theorem

Let X be a random variable with a Hypergeometric distribution $\text{HyperGeom}(N, r, n)$.
If $N, r \gg n$:

$$\text{HyperGeom}(N, r, n) \approx \text{Binomial}\left(n, p = \frac{r}{N}\right)$$

Note that when comparing a Hypergeometric distribution to its Binomial approximation, the expected values are the same; however, the exact variance is always less than or equal to the approximation due to the extra correction factor.

Example

Select 5 balls at random from an urn containing 300 red balls and 200 blue balls. Let $X =$ the number of selected red balls.

$$X \sim \text{HyperGeom}(500, 200, 5)$$

$$P(X = 3) = \frac{\binom{300}{3}\binom{500-300}{5-3}}{\binom{500}{5}} = \frac{\binom{300}{3}\binom{200}{2}}{\binom{500}{5}} = \frac{4455100 \cdot 19900}{255244687600} = 0.3473$$

$$X \sim \text{HyperGeom}(5, 0.6)$$

$$P(X = 3) \approx \binom{5}{3}(0.6)^3(0.4)^2 = 10 \cdot 0.216 \cdot 0.16 = 0.3456$$

$$E(X) = np = 5 \cdot 0.6 = 3$$

$$V(X) = np(1-p)\frac{N-n}{N-1} = 5 \cdot 0.6 \cdot 0.4 \cdot \frac{495}{499} = 1.1904$$

$$\sigma = \sqrt{1.1904} \approx 1.0910$$