

Hypergeometric Distribution

Data Science - Explanation Sheet

Definition

The hypergeometric distribution is a discrete probability distribution that gives the probability of obtaining a specific number of "successes" in a fixed number of draws from a finite population, **without replacement**, where the population is divided into two groups: successes and failures.

Why It Matters

It models situations where each draw affects the probability of the next draw, since items are not replaced after selection. This makes it different from the binomial distribution, which assumes independent trials with replacement.

Formula

$$P(X = k) = [C(K, k) \times C(N-K, n-k)] / C(N, n)$$

N = total population size | K = number of success items in population | n = number of draws (sample size) | k = number of successes observed | C(a, b) = combination (a choose b)

Conditions for Use

1. Population size (N) is finite and fixed.
2. Sampling is done without replacement.
3. There are exactly two outcomes: success or failure.
4. Probability of success changes with each draw.

Worked Example

A box has 10 balls: 4 red and 6 blue. If 3 balls are drawn without replacement, what is the probability of getting exactly 2 red balls?

N = 10, K = 4, n = 3, k = 2

$$P(X=2) = [C(4,2) \times C(6,1)] / C(10,3) = (6 \times 6) / 120 = 36/120 = \mathbf{0.30}$$

There is a 30% chance of drawing exactly 2 red balls.

Applications in Data Science

- Quality control / defect sampling in a batch of products
- A/B testing and survey sampling
- Feature selection problems involving finite sets
- Anomaly detection in small, fixed datasets

Hypergeometric vs. Binomial Distribution

Aspect	Hypergeometric	Binomial
Replacement	No	Yes
Trials	Dependent	Independent
Population	Finite	Infinite (or very large)