What is an introduction to the Hypergeometric Distribution?

Authored by

stats writer

December 26, 2025

The Hypergeometric Distribution is a fundamental concept within the field of probability distribution, utilized specifically when analyzing samples drawn from a finite population without replacement. Unlike the Binomial Distribution, which assumes independence and replacement, the Hypergeometric Distribution models situations where the outcome of one draw affects the subsequent draws, reflecting a change in the underlying population composition. This distribution allows us to calculate the exact probability of observing a specified number of successes in a sample when the sampling process ensures that an item cannot be selected more than once.

It is specifically designed to answer questions like: If you have a finite group of items, some of which possess a specific trait, and you randomly pull a smaller subset from that group, what is the chance that your subset contains exactly a certain number of items with the desired trait? This distinction makes it invaluable in quality control, population estimation, and contexts like card games, where the principle of sampling without replacement is inherent to the process.

Understanding Sampling Without Replacement

The hypergeometric distribution provides the mathematical framework for calculating probabilities when we select items sequentially from a fixed, finite pool, and those selected items are not returned to the pool. This scenario, known as sampling without replacement, ensures that the draws are dependent events. This dependency is the key feature distinguishing it from other probability models, such as the Binomial Distribution, which assumes that trials are independent.

Formally, the hypergeometric distribution describes the probability of choosing exactly k objects exhibiting a certain feature--designated as "successes"--in n total draws. These draws are made from a finite population of size N that originally contains K objects possessing that specific feature. Understanding these four core parameters--$N, K, n,$ and k--is the first crucial step in mastering this powerful statistical tool.

The Mathematical Foundation of the Hypergeometric Distribution

When a random variable X follows a hypergeometric distribution, the probability of obtaining exactly k successes, denoted $P(X=k)$, is determined by calculating the ratio of favorable outcomes to the total possible outcomes. Both the numerator and the denominator rely heavily on principles of combinatorics, specifically the mathematical concept of combination, often written as $C(n, k)$ or $\binom{n}{k}$.

The formula combines three essential components: the number of ways to select k successes from the total K successes available in the population, multiplied by the number of ways to select the remaining $(n-k)$ failures from the available $(N-K)$ failures, all divided by the total number of ways to select a sample of size n from the entire population N. The resulting probability function is expressed as follows:

P(X=k) = KCk (N-KCn-k) / NCn

Decoding the Hypergeometric Formula P(X=k)

To ensure clarity when applying this formula, it is essential to define each parameter and its specific role in calculating the final probability. Each variable represents a count of objects or trials, providing the necessary input for the combination calculations, which measure how many ways selections can be made without regard to order:

N: Represents the total size of the population size. This is the total number of items available for selection before any draws occur.
K: Represents the total number of items in the population that possess the specific feature or characteristic defined as a "success."
n: Represents the chosen sample size. This is the number of items drawn from the population.
k: Represents the number of items in the sample (the number of successes) that we are interested in calculating the probability for. It must satisfy $0 le k le n$ and $0 le k le K$.
KCk: This term calculates the number of ways to choose k successes from the K available successes in the population.

Practical Example: Calculating Card Probabilities

A classic illustration of the Hypergeometric Distribution involves drawing cards from a standard deck. Consider a standard deck of 52 cards, which contains 4 Queens. Suppose we randomly select a card, and then, without replacing the first card, we randomly select a second card. We want to determine the probability that both cards drawn are Queens.

This problem perfectly fits the hypergeometric model because we are sampling from a finite population (52 cards) and operating without replacement, meaning the probability of the second draw is dependent on the result of the first. We must carefully establish the parameters based on the specific constraints of the problem:

N (Population Size): 52 cards (Total cards in the deck)
K (Population Successes): 4 Queens (Total Queens available)
n (Sample Size): 2 draws (Total cards we pick)
k (Sample Successes): 2 Queens (Target number of Queens in our sample)

By substituting these values into the formula, we perform the necessary combination calculations:

P(X=2) = 4C2 (52-4C2-2) / 52C2

Calculating the resulting combinations:

4C2 (Ways to choose 2 Queens from 4) = 6
48C0 (Ways to choose 0 non-Queens from 48) = 1
52C2 (Total ways to choose 2 cards from 52) = 1326

Plugging these resulting numbers back into the probability function yields:

P(X=2) = (6 * 1) / 1326 = 6 / 1326 ≈ **0.00452**.

The resulting probability is approximately 0.452%. This low figure intuitively makes sense; the likelihood of drawing two specific cards sequentially without replacement is quite small, confirming the precision of the hypergeometric model.

Key Properties of the Distribution (Mean and Variance)

Beyond calculating specific probabilities, the Hypergeometric Distribution possesses important statistical properties, specifically its expected value (mean) and its variance. These measures are crucial for understanding the central tendency and the degree of dispersion in the distribution when sampling repeatedly.

The **Mean** (E or mu) represents the average number of successes expected in the sample n. It is calculated by multiplying the sample size (n) by the proportion of successes in the total population (K/N):

The mean of the distribution is: **(nK) / N**

The **Variance** ($text{Var}$ or $sigma^2$) measures the spread or variability of the distribution around the mean. The formula for the variance of the hypergeometric distribution is complex due to the inclusion of the "finite population correction factor" ($(N-n)/(N-1)$), which accounts for the reduction in variability that occurs because the draws are dependent (i.e., sampling without replacement).

The variance of the distribution is: **(nK)(N-K)(N-n) / (N2(N-1))**

Practice Problem Set 1: The Four-Card Draw

To further solidify your understanding of applying the hypergeometric distribution, we will work through several typical practice scenarios. These problems require careful identification of the four key parameters (N, K, n, k) before calculating the probability.

Problem 1: Suppose we randomly pick four cards from a standard 52-card deck without replacement. What is the probability that exactly two of the four selected cards are Queens?

We begin by defining the parameters based on the question:

N: population size = 52 cards
K: number of objects in population with a certain feature = 4 Queens
n: sample size = 4 draws
k: number of objects in sample with a certain feature = 2 Queens

We substitute these values into the hypergeometric formula, calculating the ways to select 2 Queens from 4, and 2 non-Queens from the remaining 48, divided by the total ways to select 4 cards from 52. Using calculation tools, we find the probability $P(X=2)$ to be approximately **0.025**. This means there is a 2.5% chance of selecting exactly two Queens when drawing four cards without replacement.

Practice Problem Set 2: Analyzing Urn Models

The hypergeometric distribution is frequently demonstrated using the classical "urn model," which provides a clean analogy for finite populations and sampling without replacement.

Problem 2: An urn contains 3 red balls and 5 green balls. You randomly choose 4 balls from the urn. What is the probability that your sample contains exactly 2 red balls?

In this scenario, we must first determine the total population size and then clearly identify the success population (red balls):

N: population size = 8 balls (3 red + 5 green)
K: number of objects in population with a certain feature = 3 red balls
n: sample size = 4 draws
k: number of objects in sample with a certain feature = 2 red balls

Applying the formula, we calculate $P(X=2)$:
$P(X=2) = frac{ binom{3}{2} binom{5}{2} }{ binom{8}{4} }$.

The calculation yields a probability of approximately **0.42857**. This result indicates a significant chance (over 42%) of selecting exactly two red balls in a sample of four, reflecting the relative proportion of red balls in the small population.

Practice Problem Set 3: Marbles and Combinatorics

Our final practice problem involves a slightly different set of parameters, continuing to emphasize the critical role of finite population size and the nature of dependent draws.

Problem 3: A basket contains 7 purple marbles and 3 pink marbles. You randomly choose 6

marbles from the basket. What is the probability that you choose exactly 3 pink marbles?

We define the parameters, treating the pink marbles as our success group:

N: population size = 10 marbles (7 purple + 3 pink)
K: number of objects in population with a certain feature = 3 pink marbles
n: sample size = 6 draws
k: number of objects in sample with a certain feature = 3 pink marbles

In this specific case, since the sample size ($n=6$) is large relative to the population ($N=10$), and our target number of successes ($k=3$) is the maximum available successes ($K=3$), the resulting probability will reflect a very constrained set of possibilities.

The probability $P(X=3)$ is calculated as:
$P(X=3) = frac{ binom{3}{3} binom{7}{3} }{ binom{10}{6} }$.

The calculation yields a probability of approximately **0.16667**, or exactly one-sixth. These practice problems illustrate the robust utility of the hypergeometric distribution across various contexts requiring precise probability calculations for dependent events.