3.4 Hypergeometric

Unit 3: Discrete Random Variables

Adapted from Blitzstein-Hwang Chapter 3.

If we have an urn filled with w white and b black balls, then drawing n balls out of the urn *with replacement* yields a $\mathrm{Bin}(n, w/(w+b))$ distribution for the number of white balls obtained in n trials, since the draws are independent Bernoulli trials, each with probability $w/(w+b)$ of success. If we instead sample *without replacement*, as illustrated in Figure 3.4.1, then the number of white balls follows a *Hypergeometric distribution*.

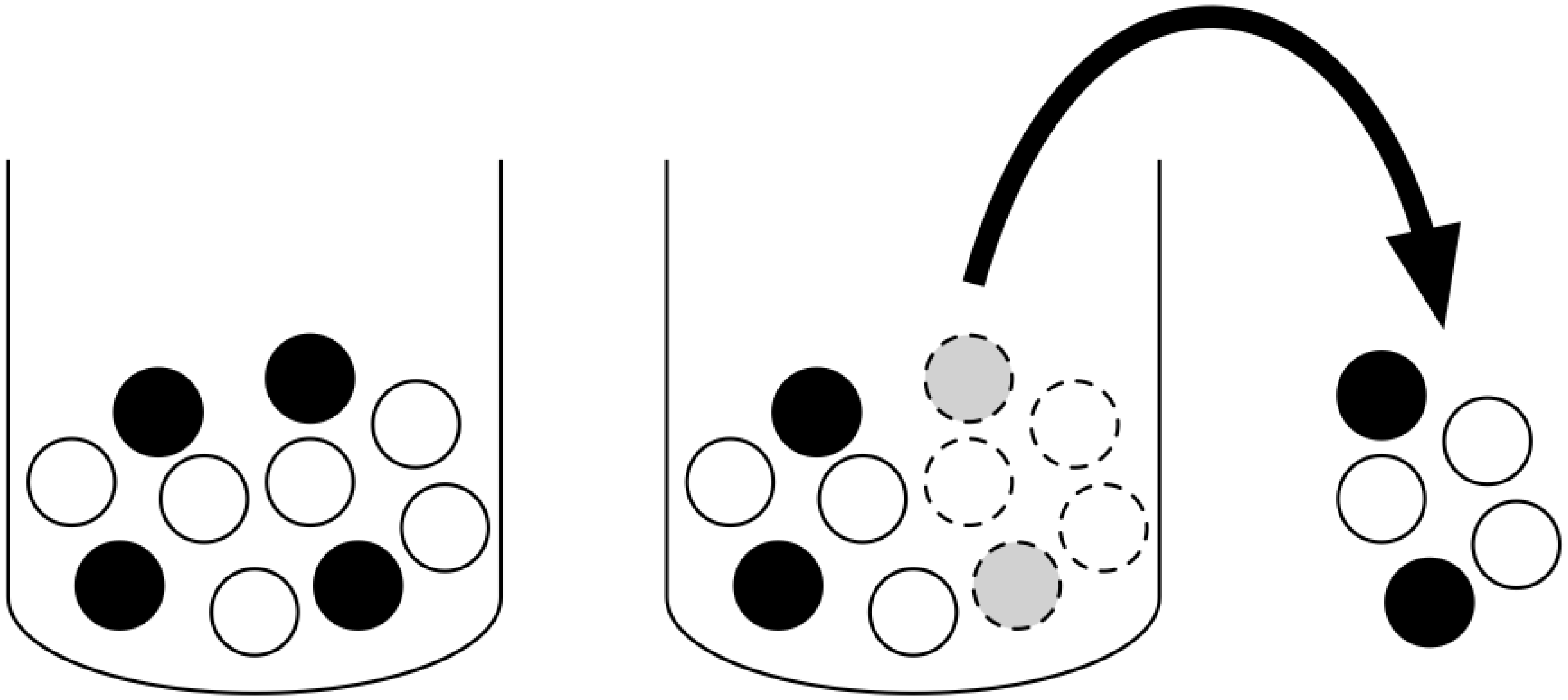

Figure 3.4.1: Hypergeometric story. An urn contains $w=6$ white balls and $b=4$ black balls. We sample $n=5$ without replacement. The number X of white balls in the sample is Hypergeometric; here we observe $X=3$.

View Larger Image
Image Description

Story 3.4.2 (Hypergeometric distribution).

Consider an urn with w white balls and b black balls. We draw n balls out of the urn at random without replacement, such that all $\binom{w+b}{n}$ samples are equally likely. Let X be the number of white balls in the sample. Then X is said to have the *Hypergeometric distribution* with parameters w, b, and n; we denote this by $X \sim \mathrm{HGeom}(w, b, n)$. As with the Binomial distribution, we can obtain the PMF of the Hypergeometric distribution from the story.

> **Theorem 3.4.3 (Hypergeometric PMF).**
>
> If $X \sim \mathrm{HGeom}(w, b, n)$, then the PMF of X is
>
> $$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$$
>
> for integers k satisfying $0 \leq k \leq w$ and $0 \leq n - k \leq b$, and $P(X = k) = 0$ otherwise.

Proof

To get $P(X = k)$, we first count the number of possible ways to draw exactly k white balls and $n - k$ black balls from the urn (without distinguishing between different orderings for getting the same set of balls). If $k > w$ or $n - k > b$, then the draw is impossible. Otherwise, there are $\binom{w}{k}\binom{b}{n-k}$ ways to draw k white and $n - k$ black balls by the multiplication rule, and there are $\binom{w+b}{n}$ total ways to draw n balls. Since all samples are equally likely, the naive definition of probability gives

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

for integers k satisfying $0 \leq k \leq w$ and $0 \leq n - k \leq b$. This PMF is valid because the numerator, summed over all k, equals $\binom{w+b}{n}$ by Vandermonde's identity, so the PMF sums to 1. The Hypergeometric distribution comes up in many scenarios which, on the surface, have little in common with white and black balls in an urn. The essential structure of the Hypergeometric story is that items in a population are classified using two sets of *tags*: in the urn story, each ball is either white or black (this is the first set of tags), and each ball is either sampled or not sampled (this is the second set of tags). Furthermore, at least one of these sets of tags is assigned completely at random (in the urn story, the balls are sampled randomly, with all sets of the correct size equally likely). Then $X \sim \mathrm{HGeom}(w, b, n)$ represents the number of twice-tagged items: in the urn story, balls that are *both* white and sampled.

Example 3.4.4 (Elk capture-recapture).

A forest has N elk. Today, m of the elk are captured, tagged, and released into the wild. At a later date, n elk are recaptured at random. Assume that the recaptured elk are equally likely to be any set of n of the elk, e.g., an elk that has been captured does not learn how to avoid being captured again. By the story of the Hypergeometric, the number of tagged elk in the recaptured sample has the $\mathrm{HGeom}(m, N - m, n)$ distribution. The m tagged elk in this story correspond to the white balls and the $N - m$ untagged elk correspond to the black balls. Instead of sampling n balls from the urn, we recapture n elk from the forest.

☣ **Warning 3.4.5 (Binomial vs. Hypergeometric).**

The Binomial and Hypergeometric distributions are often confused. Both are discrete distributions taking on integer values between 0 and n for some n, and both can be interpreted as the number of successes in n Bernoulli trials (for the Hypergeometric, each tagged elk in the recaptured sample can be considered a success and each untagged elk a failure). However, a crucial part of the Binomial story is that the Bernoulli trials involved are *independent*. The Bernoulli trials in the Hypergeometric story are *dependent*, since the sampling is done without replacement: knowing that one elk in our sample is tagged decreases the probability that the second elk will also be tagged.