Selecting Balls from urns with partial replacement rules

Enrique Treviño

LAKE FOREST COLLEGE

ISMAA 2025
March 28, 2025

Polymath REU

- A group of mathematicians from across the country created Polymath REU in 2020: Kira Adaricheva, Ben Brubaker, Pat Devlin, Steven J. Miller, Vic Reiner, Alexandra Seceleanu, Adam Sheffer, Yunus Zeytuncu
- The idea is to have professors mentor students from across the globe. To be accepted into the program, students needed to show they had taken one upper level class and have one letter recommendation. Most students accepted (in 2020 it was 300 out of 350).
- Each professor gets between 15 and 25 students to mentor on a research project. Students can contribute as much as they can, or as little as they want, but the opportunity is presented to them.

My experience in 2021

- I was one of the research mentors in 2021.
- I recruited a student of mine, Maximilano Sánchez Garza, to be my research assistant.
- 8 week program.
- I led a probability research project with the following students:
  - Julian Burden (Gettysburg)
  - Chandramauli Chakraborty (IIT, India)
  - Qizhou Fang (UC Irvine)
  - Lisa Lin (Rice)
  - Nasser Malibari (UNSW Sydney, Australia)
  - Sammi Matoush (Washington University in St. Louis)
  - Isaiah Milbank (U. Minnesota)
  - Zahan Parekh (Yale)
  - Martín Prado Guerra (Universidad de los Andes, Colombia)
  - Rachael Ren (U. Washington)
  - Qizhao Rong (Cuny)
  - Eli Sun (Rochester)
  - Daisuke Yamada (Carleton)

Our project

- Suppose you have an urn with R red balls and W white balls.
- In probability there are two classic scenarios to sample n balls from the urn.
  - When you take out a ball, you record what you got and you return the ball in the urn. (Called “sampling with replacement”)
  - When you take out a ball, you record what you got and you keep it. (Called “sampling without replacement”)

Let X be the number of white balls retrieved under some sampling procedure.

- Let X_r be when we sample with replacement,

$$\mathbb{E}(X_r) = n\frac{W}{R+W}.$$

- Let X_w be when we sample without replacement,

$$\mathbb{E}(X_w) = n\frac{W}{R+W}.$$

- Note, the probability distributions for X_r and X_w are different but they have the same expected value.

Preferential Sampling

- Let X be the number of white balls when we take a sample of n with the following procedure. If we take out a white ball, we keep it (i.e., we don't replace it), and if we take a red ball, we return it (i.e., we replace). We call this sampling *preferential sampling*.
- The expected value of X will not be the same as with X_r or X_w.
- Engbers and Hammett proved

$$\mathbb{E}[X] > \frac{3}{4}\frac{nW}{R+W} = \frac{3}{4}\mathbb{E}[X_r].$$

- Useful recursion:

$$\mathbb{E}(n,R,W) = \frac{W}{N}(1+\mathbb{E}(n-1,R,W-1)) + \frac{R}{N}\mathbb{E}(n-1,R,W),$$

Questions

- Can we improve the lower bound that Engbers and Hammett proved?
- Can we find the infimum of the ratio between

$$\frac{\mathbb{E}[X]}{\mathbb{E}[X_r]}?$$

Our Results

We can improve the lower bound:

Theorem

$$\mathbb{E}[X] > \frac{4}{5}\frac{nW}{R+W} = \frac{4}{5}\mathbb{E}[X_r].$$

We can also predict what the infimum is, although we weren't able to prove it:

Conjecture

The greatest real number ℓ satisfying

$$\mathbb{E}[X] > \ell \cdot \mathbb{E}[X_r],$$

is

$$\ell = 2 - 2e^{-W(1)} \approx 0.865713,$$

where $W(t)$ is the W Lambert function.

Numerics

n	$\mathbb{E}[X]/\mathbb{E}[X_r]$
2	0.9166666666666666
10	0.8740608411049864
100	0.8665182832864797
1,000	0.8657936212024325
10,000	0.8657214365522173
100,000	0.865714220889307
conjecture	$2 - 2e^{-W(1)} \approx 0.865713$

Table: Some values of $\mathbb{E}[X]/\mathbb{E}[X_r]$ when $R = W = n$.

Where does our conjecture come from?

- Consider $\mathbb{E}[X]$, when $R = W = n$ for some fixed large value of n.
- Let X_i be a random variable denoting the number of draws between white balls, given that i have already been drawn. It is clear that we will have $n - i$ white balls and $2n - i$ balls in total, so we have
$$\mathbb{E}(X_i) = \frac{2n - i}{n - i} = 2 + \frac{i}{n - i}.$$
- Applying linearity of expectation, we get

$$\mathbb{E}(X_0 + X_1 + \ldots + X_{k-1}) = 2k + \sum_{i=1}^{k-1} \frac{i}{n - i}.$$

- Therefore, $\mathbb{E}[X]$ should be about k when $2k + \sum_{i=1}^{k-1} \frac{i}{n-i} \approx n$. For sufficiently large n, we can approximate this sum as

$$n \approx 2k + \sum_{i=1}^{k-1} \frac{i}{n-i} \approx 2k + \int_1^{k-1} \frac{t\,\mathrm{d}t}{n-t} = 2 + k + n\log\left(\frac{n-1}{n+1-k}\right).$$

- Let $k = \alpha n$ (where α is a constant). As $n \to \infty$, we get

$$\alpha - \log(1-\alpha) = 1.$$

- Thus, for large n, we should have $\mathbb{E}[X]/\mathbb{E}[X_r] \approx \frac{2k}{n} \approx 2\alpha$.
- While $\alpha - \log(1-\alpha) = 1$ does not have a clean solution, α can be solved using Lambert's $W(t)$, function, yielding $\alpha = 1 - e^{-W(1)}$ and $2\alpha = 2 - 2e^{-W(1)}$.

Thank you!