Selecting Balls From Urns With Partial Replacement Rules

J. Burden, C. Chakraborty, Q. Fang, L. Lin,
N. Malibari, S. Matoush, I. Milbank, Z. Parekh,
M. Prado, R. Ren, Q. Rong, M. Sánchez Garza,
E. Sun, E. Treviño, and D. Yamada

Abstract - Consider an urn with an initial state of $R \geq 1$ red balls and $W \geq 1$ white balls. Draw a ball from the urn, uniformly at random, and note its color. If the ball is white, do not replace it; if the ball is red, do replace it. Define this sampling rule to be $\mathcal{P}$, for the *Preferential distribution*. We study X, the random variable denoting the number of white balls drawn under the sampling rule $\mathcal{P}$ for a sample size n. It is known that $\mathbb{E}(X)$ is bounded below by $\frac{3}{4}\frac{nW}{N}$ and bounded above by $\frac{nW}{N}$. In this paper we improve the lower bound, give a heuristic for the best possible lower bound, and we explore some properties of a generalization of this sampling rule, we call *Super-Preferential*, where the probability of retaining a white ball is w and the probability of retaining a red ball is r.

Keywords : sampling rules; expected value

Mathematics Subject Classification (2020) **:** 60C05

1 Introduction

Consider an urn with an initial state of $R \geq 1$ red balls and $W \geq 1$ white balls. Draw a ball from the urn, uniformly at random, and note its color. If the ball is white, do not replace it; if the ball is red, do replace it. Define this sampling rule to be $\mathcal{P}$, for the *Preferential distribution*, as introduced by Engbers and Hammett [1].

Let X be a random variable denoting the number of white balls drawn under the sampling rule $\mathcal{P}$ for a sample size n. The expectation of X, written $\mathbb{E}(X) := \mathbb{E}(n, R, W)$, can be obtained recursively as a function of the initial state with R red balls, W white balls, and a predetermined sample size, $n \in [1, \min\{R, W\}]$, following the relation

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

where $N := R + W$ and $\mathbb{E}(0, R, W) := 0$. Of particular interest is how $\mathcal{P}$ is related to the binomial and hypergeometric distributions, both with expectations $\mathbb{E}(B) = \mathbb{E}(H) = n\frac{W}{N}$

for $B \sim \text{Bin}\left(n, \frac{W}{N}\right)$ and $H \sim \text{HyperGeom}(N, W, n)$. Engbers and Hammett [1] proved that the ratio of the expectations of X and B is bounded; in particular, they proved that

$$c(n, R, W) = \frac{\mathbb{E}(n, R, W)}{n\frac{W}{N}} \in \left(\frac{3}{4}, 1\right].$$

Next, consider an urn with the same initial starting conditions, but where a red ball is kept with a probability r and replaced with probability $1-r$, and a white ball is kept with probability w and replaced with probability $1-w$. We denote this sampling rule as $\mathcal{S}$ for the *Super-Preferential distribution* with a similarly defined random variable $X_{(r,w)}$ denoting the number of white balls drawn under $\mathcal{S}$ with retention probabilities (r, w). Related distributions can be found by setting r and w as specific values; specifically, $(0,1)$ corresponds to the preferential, $(0,0)$ to the binomial, and $(1,1)$ to the hypergeometric. Note that, in [1], r and w are the replacement probabilities, whereas here we defined it as the opposite.

Engbers and Hammett [1] pose a number of questions. Addressed here are the following:

1. Can the bounds on $c(n, R, W)$ be improved?
2. Considering the super-preferential distribution, are there closed form expressions for the density and expectation, both under specific values of r, w and in general?

Regarding the first question, we are able to improve on their lower bound. Namely, in Section 3, we prove

Theorem 1.1 $c(n, R, W) > \frac{4}{5}$.

Furthermore, in Section 4, we give a heuristic that suggests the best lower bound for $c(n, R, W)$ is $2 - 2e^{-W(1)} \approx 0.865713$, where $W(t)$ is the Lambert W function [2].

The second question is addressed in Section 5, where we give a formula for the distribution of $X_{(r,w)}$ (Theorem 5.2) and we prove some results regarding the expected value. We also conjecture $\mathbb{E}_{0,1}(n, R, W) \leq \mathbb{E}_{r,w}(n, R, W) \leq \mathbb{E}_{1,0}(n, R, W)$, relating the super-preferential formula to the preferential formula. We can prove the conjecture if a generalization of Lemma 2.1 to the super-preferential holds, but we have been unable to prove such a generalization.

2 Technical Lemmas on $c(n, R, W)$

Recalling from the introduction, $c(n, R, W)$ expresses the ratio between the expectations of the preferential and the binomial/hypergeometric, i.e.,

$$c(n, R, W) = \frac{\mathbb{E}(n, R, W)}{n\frac{W}{N}}.$$

We proceed to understand how the function c behaves when n, R, and W vary.

We first show that $c(n, R, W)$ is weakly increasing with respect to the parameters R and W. From this, we will be able to prove a minimization for $c(n, R, W)$ from which an improved lower bound can be derived.

Lemma 2.1 *For all n, R, and W positive integers with $2 \leq n \leq \min\{R, W\}$, we have*

$$c(n, R, W) < c(n, R, W+1). \tag{2}$$

Proof. Appealing to the definition of $c(n, R, W)$, we can rewrite inequality (2) as

$$\mathbb{E}(n, R, W) < \frac{W(N+1)}{(W+1)N}\mathbb{E}(n, R, W+1), \tag{3}$$

which we proceed to prove by induction over n.

For the base case, we want to show that

$$\mathbb{E}(2, R, W) < \frac{W(N+1)}{(W+1)N}\mathbb{E}(2, R, W+1). \tag{4}$$

From [1], we know that

$$\mathbb{E}(2, R, W) = \frac{W}{N}\left(1 + \frac{R}{N} + \frac{W-1}{N-1}\right). \tag{5}$$

For $W+1$ instead of W, this yields $\mathbb{E}(n, R, W+1) = \frac{W+1}{N+1}\left(1 + \frac{R}{N+1} + \frac{W}{N}\right)$. Substituting back into (4), we want to show

$$\frac{W}{N}\left(1 + \frac{R}{N} + \frac{W-1}{N-1}\right) < \frac{W}{N}\left(1 + \frac{R}{N+1} + \frac{W}{N}\right).$$

Cancelling W/N, subtracting 1, and using that $N = R + W$, this is equivalent to

$$\begin{aligned}
\frac{R}{N} + \frac{W-1}{N-1} &< \frac{R}{N+1} + \frac{W}{N} \\
\frac{R}{N} - \frac{R}{N+1} &< \frac{W}{N} - \frac{W-1}{N-1} \\
\frac{R}{N(N+1)} &< \frac{R}{N(N-1)} \\
\frac{1}{N+1} &< \frac{1}{N-1}.
\end{aligned}$$

This inequality clearly holds for all positive integers $R, W \geq 2$, completing the base case.

Now, for the inductive hypothesis, assume that, for some $n \geq 2$ and for every $R, W \geq n$, inequality (3) holds. We seek to prove

$$\mathbb{E}(n+1, R, W) < \frac{W(N+1)}{(W+1)N}\mathbb{E}(n+1, R, W+1) \tag{6}$$

for any $R, W \geq n+1$.

From the recursion, we have

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

and

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

Substituting these equations back into inequality (6) and simplifying, we obtain

$$W\mathbb{E}(n, R, W-1) + (R-W)\mathbb{E}(n, R, W) < \frac{WR}{W+1}\mathbb{E}(n, R, W+1), \tag{7}$$

which is therefore equivalent to our initial inequality. From the inductive hypothesis, we have

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

and

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

These can be expressed as

$$W\mathbb{E}(n, R, W-1) < \frac{(W-1)N}{N-1}\mathbb{E}(n, R, W) \tag{8}$$

and

$$\frac{RN}{N+1}\mathbb{E}(n, R, W) < \frac{WR}{W+1}\mathbb{E}(n, R, W+1). \tag{9}$$

We claim that

$$\frac{(W-1)N}{N-1}\mathbb{E}(n, R, W) \leq \frac{RN}{N+1}\mathbb{E}(n, R, W) - (R-W)\mathbb{E}(n, R, W). \tag{10}$$

If this last inequality holds, then from (8), (9), and (10) it follows that

$$W\mathbb{E}(n, R, W-1) + (R-W)\mathbb{E}(n, R, W) \leq \frac{RN}{N+1}\mathbb{E}(n, R, W) < \frac{WR}{W+1}\mathbb{E}(n, R, W+1),$$

that is exactly inequality (7).

To prove (10), notice that it can be written as

$$0 \leq \left(\frac{2R}{(N+1)(N-1)}\right)\mathbb{E}(n, R, W),$$

which is true for all $R, W \geq 2$, thus proving the claim, which completes the induction and finishes the proof. $\square$

Lemma 2.2 *For all n, R, and W positive integers with $2 \leq n \leq \min\{R, W\}$ and $R+1 \geq W$, we have*

$$c(n, R, W) \leq c(n, R+1, W),$$

where the equality holds if and only if $n = 2$ and $R + 1 = W > 2$.

Proof. Like in the proof of Lemma 2.1, we express this inequality in terms of the expectation, obtaining

$$\mathbb{E}(n, R, W) \leq \frac{N+1}{N}\mathbb{E}(n, R+1, W), \tag{11}$$

which we proceed to prove by induction over n.

For the base case, using (5) we want to show

$$\frac{W}{N}\left(1 + \frac{R}{N} + \frac{W-1}{N-1}\right) \leq \frac{W}{N}\left(1 + \frac{R+1}{N+1} + \frac{W-1}{N}\right),$$

which reduces to

$$\begin{aligned}
\frac{R}{N} + \frac{W-1}{N-1} &\leq \frac{R+1}{N+1} + \frac{W-1}{N} \\
\frac{W-1}{N-1} - \frac{W-1}{N} &\leq \frac{R+1}{N+1} - \frac{R}{N} \\
\frac{W-1}{N(N-1)} &\leq \frac{W}{N(N+1)} \\
(W-1)(N+1) &\leq W(N-1) \\
W &\leq R+1.
\end{aligned}$$

Since $R + 1 \geq W$, the $n = 2$ case is settled.

For the inductive hypothesis, assume that, for some $n \geq 2$ and for every $R, W \geq n$ such that $R + 1 \geq W$, inequality (11) holds. We want to show

$$\mathbb{E}(n+1, R, W) < \frac{N+1}{N}\mathbb{E}(n+1, R+1, W) \tag{12}$$

for $R, W \geq n + 1$ and $R + 1 \geq W$.

Note that for $R + 1 \geq W$,

$$N - 1 - \frac{(R - W + 1)}{N} \leq N - 1. \tag{13}$$

Therefore, (13) implies

$$\begin{aligned}
N(N-1) - (R - W + 1) &\leq N(N-1) \\
N^2R + N(W-1) - (R - W + 1) &\leq N^2R + N(N-1) - RN \\
(N-1)(N+1)R + (W-1)(N+1) &\leq (R+1)N(N-1) \\
\frac{R}{N} + \frac{W-1}{N(N-1)} &\leq \frac{R+1}{N+1}.
\end{aligned} \tag{14}$$

Now, from the recursive formula for expected value, applying (3) when $W \to W-1$, and (14), we get

$$\begin{aligned}
\mathbb{E}(n+1,R,W) &= \frac{W}{N}\left(1+\mathbb{E}(n,R,W-1)\right) + \frac{R}{N}\mathbb{E}(n,R,W) \\
&= \frac{W}{N}\left(1+\frac{N-1}{N}\mathbb{E}(n,R,W-1)\right) + \frac{W}{N^2}\mathbb{E}(n,R,W-1) + \frac{R}{N}\mathbb{E}(n,R,W) \\
&< \frac{W}{N}\left(1+\frac{N-1}{N}\mathbb{E}(n,R,W-1)\right) + \left(\frac{W-1}{N(N-1)} + \frac{R}{N}\right)\mathbb{E}(n,R,W) \\
&\leq \frac{W}{N}\left(1+\frac{N-1}{N}\mathbb{E}(n,R,W-1)\right) + \frac{R+1}{N+1}\mathbb{E}(n,R,W). \qquad (15)
\end{aligned}$$

Finally, applying the inductive hypothesis, we determine that

$$\begin{aligned}
&\frac{W}{N}\left(1+\frac{N-1}{N}\mathbb{E}(n,R,W-1)\right) + \frac{R+1}{N+1}\mathbb{E}(n,R,W) \\
&\leq \frac{W}{N}\left(1+\mathbb{E}(n,R+1,W-1)\right) + \frac{R+1}{N}\mathbb{E}(n,R+1,W) \\
&= \frac{N+1}{N}\left(\frac{W}{N+1}\left(1+\mathbb{E}(n,R+1,W-1)\right) + \frac{R+1}{N+1}\mathbb{E}(n,R+1,W)\right) \\
&= \frac{N+1}{N}\mathbb{E}(n+1,R+1,W). \qquad (16)
\end{aligned}$$

Therefore, combining (15) and (16) yields (12), as desired. The equality case of (11) follows from the analysis done in the case $n = 2$ and the inequality proved in the inductive step. □

Now, we show that c is decreasing with respect to n.

Lemma 2.3 *For all n, R, W with $1 \leq n \leq \min\{R, W\} - 1$, we have*

$$c(n,R,W) \geq c(n+1,R,W).$$

Proof. By the definition of c, the result is equivalent to

$$\mathbb{E}(n+1,R,W) \leq \frac{n+1}{n}\mathbb{E}(n,R,W), \qquad (17)$$

which we proceed to prove by induction over n. For the base case $n = 1$, from [1, eq.

(10)], we have that

$$\begin{aligned}
\mathbb{E}(2,R,W) &= \frac{W}{N}\left(1+\frac{R}{N}+\frac{W-1}{N-1}\right)\\
&< \frac{W}{N}\left(1+\frac{R}{N-1}+\frac{W-1}{N-1}\right)\\
&= \frac{W}{N}\left(1+\frac{N-1}{N-1}\right)\\
&= \frac{2W}{N}\\
&= 2\mathbb{E}(1,R,W). \qquad (18)
\end{aligned}$$

Now, assume the result for $1,2,\ldots,n-1$, and we will show it for n. From (1) and the induction hypothesis, we have

$$\begin{aligned}
\mathbb{E}(n+1,R,W) &= \frac{W}{N}\left(1+\mathbb{E}(n,R,W-1)\right)+\frac{R}{N}\mathbb{E}(n,R,W)\\
&\leq \frac{W}{N}\left(1+\frac{n}{n-1}\mathbb{E}(n-1,R,W-1)\right)+\frac{R}{N}\left(\frac{n}{n-1}\mathbb{E}(n-1,R,W)\right). \qquad (19)
\end{aligned}$$

On the other hand, again by (1), we have

$$\frac{n+1}{n}\mathbb{E}(n,R,W)=\left(\frac{n+1}{n}\right)\left(\frac{W}{N}\left(1+\mathbb{E}(n-1,R,W-1)\right)+\frac{R}{N}\mathbb{E}(n-1,R,W)\right). \qquad (20)$$

So, to show (17), it is enough to show that the right hand side of (19) is less than or equal to the right hand side of (20).

Multiplying by nN and rearranging terms, it reduces to proving

$$\frac{W}{n-1}\mathbb{E}(n-1,R,W-1)+\frac{R}{n-1}\mathbb{E}(n-1,R,W)\leq W. \qquad (21)$$

From [1, Thm. 2], we know that $\mathbb{E}(n-1,R,W-1)\leq(n-1)\frac{W-1}{N-1}$ and $\mathbb{E}(n-1,R,W)\leq(n-1)\frac{W}{N}$. From here it follows that

$$\begin{aligned}
\frac{W}{n-1}\mathbb{E}(n-1,R,W-1)+\frac{R}{n-1}\mathbb{E}(n-1,R,W) &\leq \frac{W(W-1)}{N-1}+\frac{RW}{N}\\
&= W\left(\frac{W-1}{N-1}+\frac{R}{N}\right)<W,
\end{aligned}$$

where the last inequality follows from the computation done in (18). Thus (21) is true, and the result follows. □

3 Improved Lower Bound for $c(n, R, W)$

Combining Lemmas 2.1 and 2.2, we see that $c(n, R, W)$ is increasing with respect to both R and W. This combination leads directly to the following lemma.

Lemma 3.1 *For all $R, W \geq n \geq 1$, we have*

$$c(n, n, n) \leq c(n, R, W),$$

where the equality holds if and only if $n = 1$ or $R = W = n$.

Proof. Suppose $R = n + a$ and $W = n + b$ for non-negative integers a, b. Apply Lemma 2.2 a times to get $c(n, n, n) \leq c(n, R, n)$, then apply Lemma 2.1 b times to get $c(n, R, n) \leq c(n, R, W)$. □

Using Lemma 3.1, we can prove a better lower bound for $c(n, R, W)$ than $\frac{3}{4}$, i.e., Theorem 1.1.

Proof. [Proof of Theorem 1.1] From Lemma 3.1, we have $c(n, n, n) \leq c(n, R, W)$, so the problem reduces to showing $c(n, n, n) > 4/5$. In terms of expectations, the desired inequality can be represented as

$$\mathbb{E}(n, n, n) > \frac{2n}{5}.$$

From here we proceed by induction.

For the base case we simply evaluate (5) at $n = R = W = 2$, thus obtaining $\mathbb{E}(2, 2, 2) = \frac{11}{12} > \frac{4}{5}$.

Now, for the induction hypothesis, we assume for some $n \geq 2$ that $\mathbb{E}(n, n, n) > \frac{2n}{5}$. Applying the recursion formula for $\mathbb{E}(n + 1, n + 1, n + 1)$, we find

$$\mathbb{E}(n + 1, n + 1, n + 1) = \frac{1}{2}(1 + \mathbb{E}(n, n + 1, n) + \mathbb{E}(n, n + 1, n + 1)). \tag{22}$$

Substituting $R = W = n$ in (11), we get

$$\frac{2n}{2n + 1}\mathbb{E}(n, n, n) \leq \mathbb{E}(n, n + 1, n).$$

Combining this last inequality with (22) we obtain that

$$\mathbb{E}(n + 1, n + 1, n + 1) \geq \frac{1}{2}\left(1 + \frac{2n}{2n + 1}\mathbb{E}(n, n, n) + \mathbb{E}(n, n + 1, n + 1)\right).$$

Then, it suffices to prove that

$$\frac{1}{2}\left(1 + \frac{2n}{2n + 1}\mathbb{E}(n, n, n) + \mathbb{E}(n, n + 1, n + 1)\right) > \frac{2(n + 1)}{5}. \tag{23}$$

From (3) and (11) we know that $\mathbb{E}(n, n+1, n+1) \geq \mathbb{E}(n, n, n)$. Using these inequalities and the induction hypothesis, we arrive to

$$\frac{1}{2}\left(1+\frac{2n}{2n+1}\mathbb{E}(n,n,n)+\mathbb{E}(n,n+1,n+1)\right) \geq \frac{1}{2}\left(1+\frac{2n}{2n+1}\mathbb{E}(n,n,n)+\mathbb{E}(n,n,n)\right)$$
$$> \frac{1}{2}\left(1+\frac{2n}{2n+1}\left(\frac{2n}{5}\right)+\frac{2n}{5}\right) = \frac{8n^2+12n+5}{20n+10} > \frac{8n^2+12n+4}{20n+10} = \frac{2(n+1)}{5}.$$

□

4 Heuristic for the Best Lower Bound on $c(n, R, W)$

Numerical evidence suggests that $c(n, n, n)$ is decreasing (see Table 1).

n	$c(n, n, n)$
2	0.9166666666666666
10	0.8740608411049864
100	0.8665182832864797
1,000	0.8657936212024325
10,000	0.8657214365522173
100,000	0.865714220889307

Table 1: Some values of $c(n, n, n)$.

Assuming that $c(n, n, n)$ is decreasing, by Theorem 1.1, $c(n, n, n)$ converges to a value as $n \to \infty$. Furthermore, from Lemma 3.1, it follows that the best lower bound for $c(n, R, W)$ is the limit as $n \to \infty$ of $c(n, n, n)$.

Our conjecture is

Conjecture 4.1 The function $c(n, n, n)$ is decreasing and

$$\lim_{n\to\infty} c(n, n, n) = 2 - 2e^{-W(1)},$$

where W is the Lambert's W function.

Note that $2 - 2e^{-W(1)} \approx 0.865713$, which is consistent with the numerical values in Table 1.

4.1 Heuristic Suggesting Conjecture 4.1

Consider $\mathbb{E}(n, n, n)$, 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

$$\mathbb{E}(X_i) = \frac{2n-i}{n-i} = 2 + \frac{i}{n-i}.$$

Then 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, $\mathbb{E}(n, n, n)$ 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

$$n \approx 2k + \sum_{i=1}^{k-1} \frac{i}{n-i} \approx 2k + \int_1^{k-1} \frac{t\,\mathrm{d}t}{n-t} = 2 + k + n \log\left(\frac{n-1}{n+1-k}\right). \tag{24}$$

Let $k = \alpha n$ (where α is a constant) and divide (24) by n. As $n \to \infty$, we get

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

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

5 Super-Preferential

A number of advancements have been made with regards to the expectation and density of the super-preferential.

5.1 General Expectation

The recurrence (1) can be generalized to the super-preferential case. This is, let $X_{(r,w)}$ be a random variable denoting the number of white balls in a sample of size n drawn, at random, from an urn initially containing R red balls and W white balls, for a total of $N = N$ total starting balls where, if the ball is red, we keep it with probability $r \in [0, 1]$ and, if it is white, we keep it with probability $w \in [0, 1]$. We denote the expected value as $\mathbb{E}\left(X_{(r,w)}\right) =: \mathbb{E}_{r,w}(n, R, W)$. Then, from the arguments used to justify (1), we can conclude that

$$\begin{aligned}\mathbb{E}_{r,w}(n, R, W) = \frac{W}{N}(w)(1 + \mathbb{E}_{r,w}(n-1, R, W-1)) + \frac{W}{N}(1-w)(1 + \mathbb{E}_{r,w}(n-1, R, W)) \\ + \frac{R}{N}(r)\mathbb{E}_{r,w}(n-1, R-1, W) + \frac{R}{N}(1-r)\mathbb{E}_{r,w}(n-1, R, W). \end{aligned} \tag{25}$$

with the initial condition $\mathbb{E}_{r,w}(0, R, W) = 0$. This generalized recurrence is in agreement with those of the binomial, hypergeometric, and preferential.

5.2 The Case of $r = w$

Consider the case where $r = w = y \in [0, 1]$. Since the probability of keeping a ball of either color is the same, it is reasonable to think that the expected value should be the same for every y. Turns out this is true and, in fact, can be shown to have the same expectation as both the binomial and hypergeometric cases.

Theorem 5.1 *If $y \in [0, 1]$, then for $n \geq 1$ we have*

$$\mathbb{E}_{y,y}(n, R, W) = n\frac{W}{N}. \tag{26}$$

Proof. We proceed by induction over n. For the base case, from (25) it is easy to obtain $\mathbb{E}_{y,y}(1, R, W) = \frac{W}{N}$. Now, assume (26) is true. Then, from (25) we have

$$\begin{aligned}
&\mathbb{E}_{y,y}(n+1, R, W)\\
&= y\frac{W}{N}\left(1 + n\left(\frac{W-1}{N-1}\right)\right) + (1-y)\frac{W}{N}\left(1 + n\frac{W}{N}\right) + ny\frac{R}{N}\left(\frac{W}{N-1}\right) + n(1-y)\frac{R}{N}\frac{W}{N}\\
&= \frac{W}{N}\left(1 + ny\left(\frac{W-1}{N-1}\right) + n(1-y)\frac{W}{N} + ny\left(\frac{R}{N-1}\right) + n(1-y)\frac{R}{N}\right)\\
&= \frac{W}{N}\left(1 + ny + n(1-y)\right) = (n+1)\frac{W}{N}.
\end{aligned}$$

□

5.3 Density for $X_{(r,w)}$

We will follow the strategy for preferential density in [1], with the proper generalizations. Let each sequence of n draws be associated with some n-letter word $\mathbf{d} := d_1 d_2 \cdots d_n$ where $d_i \in \{\text{red}_1, \text{red}_2, \text{white}_1, \text{white}_2\}$, where red_1 represents balls that are to be retained and red_2 represents balls that are to be replaced. Define similarly for white_1 and white_2. Suppose $\mathbf{d}$ contains k white balls, where the locations of the white balls that will be retained are $w_1 < w_2 < \ldots < w_t$ (this means $d_{w_1}, d_{w_2}, \ldots, d_{w_t}$ are the white balls that are retained) and the locations of the red balls that will be retained are $r_1 < r_2 < \ldots < r_s$. Let $F(i)$ be the number of white balls in the first i letters and $G(i)$ be the number of red balls in the first i letters. Note that when a white ball appears between w_i and w_{i+1}, then the probability of such an occurrence is of the form x/y where $x = W - i$. This happens for $i = 0, 1, \ldots, t$ (if we define $w_0 = 0$, $w_{t+1} = n$). Something similar happens with red balls between r_j and r_{j+1} for $j = 0, 1, \ldots, s$ (if we define $r_0 = 0$, $r_{s+1} = n$). Now, if we pick any ball (red or white) between two balls that are not replaced (whether the non-replaced ones are red or white), then the probability is of the form x/y where $y = N - \ell$, where ℓ is the number of non-replacements so far. Suppose the non-replaced balls have indices $\nu_1 < \nu_2 < \ldots < \nu_{r+s}$ (define $\nu_0 = 0$ and $\nu_{r+s+1} = n$). Then

$$\mathbb{P}(\mathbf{d}) = w^t(1-w)^{k-t}r^s(1-r)^{n-k-s}\prod_{i=0}^{t}(W-i)^{F(w_{i+1})-F(w_i)} \cdot \prod_{j=0}^{s}(R-j)^{G(r_{j+1})-G(r_j)}\prod_{\ell=0}^{r+t}\left(\frac{1}{N-\ell}\right)^{\nu_{\ell+1}-\nu_\ell}. \tag{27}$$

Let $|\mathbf{d}|$ be the number of white balls in the word $\mathbf{d}$.

Theorem 5.2 *Under sampling rule $\mathcal{S}$, let $X_{(r,w)}$ denote the number of white balls in our sample of size n. Then, for each $k \in \{0, 1, 2, \ldots, n\}$ the probability mass function of $X_{(r,w)}$ is given:*

$$\mathbb{P}(X_{(r,w)} = k) = \sum_{|\mathbf{d}|=k}\mathbb{P}(\mathbf{d}),$$

where $\mathbb{P}(\mathbf{d})$ is as in (27).

r	w	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$\mathbb{E}_{r,w}(3,4,3)$
2/3	1/3	0.13955296	0.45537199	0.342727567	0.0623474787	1.32786956
1/3	2/3	0.16365403	0.474354821	0.316358925	0.0456322211	1.24396933
0.5	0.5	0.151749271	0.464820214	0.329397473	0.0540330418	1.28571429
0.8	0.5	0.129586006	0.470919339	0.345461613	0.054033042	1.32394169
0.5	0.8	0.151749271	0.487531584	0.321869776	0.0388493683	1.24781924
0.1	0.5	0.179830904	0.458157434	0.30797862	0.0540330418	1.2362138
0.5	0.1	0.151749271	0.434538387	0.339869776	0.073842566	1.33580564

Table 2: The super-preferential distribution when $n = 3, R = 4, W = 3$ for various retention probabilities, r, w.

5.4 Bounds on the Expectation of the Super-Preferential

We conjecture that the expectation of a super-preferential is bounded below by the expectation of the preferential ($r = 0$ and $w = 1$) and bounded above by the expectation of the distribution with reversed retention probabilities ($r = 1$ and $w = 0$). That is

Conjecture 5.3 Let R, W, and n be positive integers such that $2 \le n \le min\{R, W\}$. Then, for any $r, w \in [0, 1]$, we have

$$\mathbb{E}_{0,1}(n, R, W) \le \mathbb{E}_{r,w}(n, R, W) \le \mathbb{E}_{1,0}(n, R, W).$$

We can prove the conjecture if we were able to prove

$$\mathbb{E}_{r,w}(n-1, R, W-1) < \mathbb{E}_{r,w}(n-1, R, W), \tag{28}$$

which would be a generalization of Lemma 2.1 to the super-preferential distribution. Our proof assuming (28) is long and intricate, so for brevity we will exclude it from the paper.

Acknowledgments

This research was done as part of the Polymath REU 2021 project. We would like to thank the organizers of the program. We would also like to thank Steven J. Miller for useful conversations and financial support for the project. Finally, we would like to thank Peter Winkler for answering some probability questions and helping with the submission process.

References

[1] J. Engbers, A. Hammett, To replace or not to replace — that is the question, *College Math. J.*, **51** (2020), 117–123.

[2] Wikipedia contributors, Lambert W function, *Wikipedia*, available online at the UTL: `https://en.wikipedia.org/wiki/Lambert_W_function`

Julian Burden
E-mail: `julianburden9@gmail.com`

Chandramauli Chakraborty
University of Chicago
Chicago, IL 60637
E-mail: `chandramaulic@uchicago.edu`

Qizhou Fang
University of California Davis
Davis, CA 95616
E-mail: `qzfang@ucdavis.edu`

Lisa Lin
Yale University
New Haven, CT 06520
E-mail: `lisa.lin@yale.edu`

Nasser Malibari
University of New South Wales
Sydney NSW 2033, Australia
E-mail: `nasser.malibari@student.unsw.edu.au`

Sammi Matoush
E-mail: `smatoush214@gmail.com`

Isaiah Milbank
E-mail: `isaiah.milbank@gmail.com`

Zahan Parekh
E-mail: zahansparekh@gmail.com

Martín Prado
Universidad de los Andes
Bogotá, Colombia, 111711
E-mail: m.prado@uniandes.edu.co

Rachael Ren
University of Texas at Austin
Austin, TX 78712
E-mail: rren@utexas.edu

Qizhao Rong
E-mail: qizhao.rong@baruchmail.cuny.edu

Maximiliano Sánchez Garza
University of Virginia
Charlottesville, VA 22904
E-mail: zha5sb@virginia.edu

Eli Sun
University of Rochester
Rochester, NY 14642
E-mail: eli_sun@urmc.rochester.edu

Enrique Treviño
Lake Forest College
Lake Forest, IL 60045
E-mail: trevino@lakeforest.edu

Daisuke Yamada
University of Wisconsin at Madison
Madison, WI 53706
E-mail: daisuke@cs.wisc.edu

Received: August 6, 2024 **Accepted**: November 14, 2024
Communicated by Serban Raianu