On an intriguing urn probability problem

1. Introduction

A recent piece ([1]) in Quanta Magazine discussed various probability puzzles posed by the mathematician Daniel Litt on social media. Among these, there was the following problem:

Imagine that you have an urn filled with 100 *balls, some red and some green. You can't see inside; all you know is that someone determined the number of red balls by picking a number between zero and 100 from a hat. You reach into the urn and pull out a ball. It's red. If you now pull out a second ball, it is more likely to be Red or Green? (or are the two colors equally likely)?*

The answer is 'Red' and the corresponding probability is $2/3$. At first glance, without calculations and as discussed in ([1]), it may well be puzzling why the probability is equal to $2/3$. The problem does cry for interpretation, has much pedagogical merit, and has also caught attention elsewhere (e.g., [2, 3]). In a variant of the problem ([1]), with the number of red balls among the 100 determined by a symmetric binomial distribution, i.e., flipping a coin to determine each ball's colour, the probability of a choosing a red ball on the second draw is instead equal to $1/2$. In this note, we focus on the determination of such probabilities which requires to know either the exact number of red balls in the urn; or as is the case here the prior distribution of red balls in the urn. We examine interesting extensions: (i) to other prior distributions on the number of red balls in the urn, such as the Binomial prior distribution, and (ii) to sampling with replacement of the initial ball. Finally, we reveal further surprising properties with regards to the categorization of probabilistic independence, contrasting sampling with and without replacement for a Binomial prior distribution.

Expectations and variances of random variables are denoted throughout as E and V, respectively.

2. Analysis

Consider a more general version with $M \geq 2$ balls in the urn, N of which are red, $M - N$ green, with N a non-degenerate random variable taking values on $\{0, 1, \dots, M\}$. The original problem has $M = 100$ and N uniformly distributed on $\{0, 1, \dots 100\}$. Let $X_i = 1$ if the i^{th} drawn ball is red, and 0 otherwise, for $i = 1, 2$. The stated problem is to determine

$$\gamma \;=\; \mathbb{P}(X_2 = 1 | X_1 = 1)\,.$$

Calculations yield $\mathbb{P}(X_1 = 1, X_2 = 1|N) = \frac{N(N-1)}{M(M-1)}$, $\mathbb{P}(X_1 = 1|N) = \frac{N}{M}$, and lead to

$$\begin{aligned}
\gamma \;=\; \frac{\mathbb{P}\big(X_1 = 1, X_2 = 1\big)}{\mathbb{P}\big(X_1 = 1\big)} \quad &= \quad \frac{\sum_{n\geq 0}\mathbb{P}(X_1 = 1, X_2 = 1, N = n)}{\sum_{n\geq 0}\mathbb{P}(X_1 = 1, N = n)} \\
&= \quad \frac{\sum_{n\geq 0}\mathbb{P}(X_1 = 1, X_2 = 1|N = n)\,\mathbb{P}(N = n)}{\sum_{n\geq 0}\mathbb{P}(X_1 = 1|N = n)\,\mathbb{P}(N = n)} \\
&= \quad \frac{E\Big\{\mathbb{P}\big(X_1 = 1, X_2 = 1|N\big)\Big\}}{E\Big\{\mathbb{P}(X_1 = 1|N)\Big\}} \\
&= \quad \frac{\frac{1}{M(M-1)}\;E\big(N(N-1)\big)}{\frac{1}{M}\;E(N)} \\
&= \quad \frac{1}{M-1}\Big(\frac{V(N)}{E(N)} \;+\; E(N) \;-\; 1\Big)\,.
\end{aligned}$$

Remark. The evaluation of the probability of the event $\{X_2 = 1\}$ is carried out above conditional on the event $\{X_1 = 1\}$. The latter precludes cases where $N = 0$ has been observed, so that a reasonable and alternative approach is to consider at the outset distributions for N such that $\mathbb{P}(N = 0) > 0$, or to condition on $N > 0$. However, the above conditional probability evaluation of γ does take into account such considerations and is equivalent to conditioning on $N > 0$. To see this, observe that a re-evaluation of γ along such lines yields on the fourth line

$$\begin{aligned}
\gamma \quad &= \quad \frac{1}{M-1}\,\frac{\mathbb{E}(N(N-1)|N \geq 1)}{\mathbb{E}(N|N \geq 1)} \;=\; \frac{1}{M-1}\,\frac{\sum_{n\geq 1} n(n-1)\,\frac{\mathbb{P}(N=n)}{\mathbb{P}(N\geq 1)}}{\sum_{n\geq 1} n\,\frac{\mathbb{P}(N=n)}{\mathbb{P}(N\geq 1)}} \\
&= \quad \frac{\sum_{n\geq 0} n(n-1)\,\mathbb{P}(N = n)}{\sum_{n\geq 0} n\,\mathbb{P}(N = n)} \;=\; \frac{1}{M-1}\,\frac{E(N(N-1))}{E(N)}
\end{aligned}$$

which matches indeed the expression above.

The above evaluation of γ applies for all values of M and distribution of N on $\{0, 1, \dots, M\}$, such as for the following situations:

(a) (Uniform) For N uniformly distributed on $\{0, 1, \ldots, M\}$, we have $E(N) = M/2$ and $\frac{V(N)}{E(N)} = \frac{M(M+2)/12}{M/2} = \frac{M+2}{6}$; with the evaluations $E(N) = \frac{1}{M+1}\sum_{k=0}^{M} k = M/2$, $E(N^2) = \frac{1}{M+1}\sum_{k=0}^{M} k^2 = \frac{1}{M+1}\frac{(M)(M+1)(2M+1)}{6}$, and $V(N) = E(N^2) - E^2(N)$; yielding

$$\gamma = \frac{1}{M-1}\Big(\frac{M+2}{6} + \frac{M}{2} - 1\Big) = \frac{2}{3}.$$

Surprisingly, the above probability does not depend on M, the case $M = 100$ corresponding to Daniel Litt's original problem.

(b) (Binomial) For $N \sim Bin(M, \alpha)$, $0 < \alpha < 1$, we have $E(N) = M\alpha$, and $\frac{V(N)}{E(N)} = (1 - \alpha)$ which yields

$$\gamma = \frac{1}{M-1}\Big((1-\alpha) + M\alpha - 1\Big) = \alpha. \tag{1}$$

This is particularly interesting since the above is independent of M, and $\gamma \geq 1/2$ iff $\alpha \geq 1/2$ with equality iff $\alpha = 1/2$; the latter case the object of Daniel Litt's social media blog. We also point out that a similar calculation yields $\mathbb{P}(X_2 = 1|X_1 = 0) = \alpha = \mathbb{P}(X_2 = 1|X_1 = 1)$, establishing independence between the draws X_1 and X_2.

(c) Discriminating the cases $\gamma > 1/2$ and $\gamma < 1/2$ is of intrinsic interest. From the general expression for γ, we obtain

$$\gamma \geq 1/2 \Longleftrightarrow \frac{V(N)}{E(N)} + E(N) \geq \frac{M+1}{2}, \tag{2}$$

with (i) the condition $E(N) \geq \frac{M+1}{2}$ sufficient for γ to exceed $1/2$; and (ii) the condition $V(N) \geq \frac{M}{4}$ implying $\gamma \geq 1/2$ in cases where $E(N) = \frac{M}{2}$, i.e. the expected number of red and green balls in the urn are equal. In the latter situation $\gamma = 1/2$ iff $V(N) = \frac{M}{4}$ which is what occurs for the binomial $Bin(M, 1/2)$ case. Although Condition (i) is simple and does not require knowing about the variance of N, we emphasize that it is not necessary as witnessed for instance with the uniform case where $E(N) = \frac{M}{2}$ is insufficient for (i), but where $\gamma = 2/3$ is much larger than $1/2$. On the other hand, revisiting the binomial case with $\alpha = \frac{M+1}{2M}$, condition (i) is satisfied with $E(N) = \frac{M+1}{2}$ and $\gamma - \frac{1}{2} = \frac{M+1}{2} - \frac{1}{2} = \frac{1}{2M}$, which becomes arbitrarily close to 0 for large M.

3. Sampling with replacement

It is also appealing to consider the initial problem with a sampling replacement scheme, where the two draws are conditionally independent on the number N of red balls in the urn. In accordance, a reasonable model for this problem is to set $X_1, X_2|p$ independently distributed as Bernoulli(p). The problem is (again) to determine $\gamma = \mathbb{P}(X_2 = 1|X_1 = 1)$. Given a prior density for p, we have $\mathbb{P}(X_2 = 1, X_1 = 1|p) = p^2$, $\mathbb{P}(X_2 = 1, X_1 = 1) = E(p^2)$, $\mathbb{P}(X_1 = 1|p) = p$, and $\mathbb{P}(X_1 = 1) = E(p)$. We thus obtain

$$\gamma = \frac{\mathbb{P}(X_2 = 1, X_1 = 1)}{\mathbb{P}(X_1 = 1)} = \frac{E(p^2)}{E(p)}. \tag{3}$$

Interesting examples include:

(i) The uniform case with $p \sim \mathbb{U}(0,1)$ which yields $\gamma = \frac{E(p^2)}{E(p)} = \frac{1/3}{1/2} = 2/3$. Observe that this matches the sampling without replacement uniform case.

(ii) Beta distribution cases with $p \sim \mathbb{B}(a,b)$. Here, one obtains $\gamma = \frac{a+1}{a+b+1}$ with the evaluations $E(p^2) = \frac{a(a+1)}{(a+b)(a+b+1)}$ and $E(p) = \frac{a}{a+b}$. Observe here that $\gamma \geq 1/2$ if and only if $a + 1 \geq b$ with equality for $a + 1 = b$.

(iii) The discrete uniform case with p distributed as $\frac{N}{M}$ and N having a discrete uniform distribution on the integers $0, 1, \ldots, M$. Here we obtain, with the same evaluations as in example (a) of Section 2

$$\gamma = \frac{E(p^2)}{E(p)} = \frac{E(N^2)}{M\,E(N)} = \frac{M(2M+1)/6}{M^2/2} = \frac{2M+1}{3M} > \frac{2}{3},$$

which is slightly larger that the sampling without replacement case.

(iv) The Binomial case with with p distributed as $\frac{N}{M}$ and $N \sim Bin(M, \alpha)$, $0 < \alpha < 1$. As in (iii), we obtain

$$\gamma = \frac{E(N^2)}{M\,E(N)} = \frac{1}{M}\,\frac{(M\alpha)^2 + M\alpha(1-\alpha)}{M\alpha} = \frac{1 + (M-1)\alpha}{M},$$

which yields for instance $\gamma = \frac{101}{200}$ for $M = 100, \alpha = 1/2$.

Positive dependence

Observe that $\gamma > \alpha$ in the above Binomial distributed scenario, in contrast to the without replacement scenario where $\gamma = \alpha$. This translates to positive dependence between X_1 and X_2, i.e., $\mathbb{P}(X_2 = 1|X_1 = 1) > \mathbb{P}(X_2 = 1)$, which makes intuitive sense. This relationship holds more generally as long as the number N of red balls is purely random, i.e., not fixed or degenerate. To see this, reconsider $\mathbb{P}(X_2 = 1|X_1 = 1)$ in (3), together with $\mathbb{P}(X_2 = 1)$ which is equal to $\mathbb{P}(X_2 = 1, X_1 = 1) + \mathbb{P}(X_2 = 1, X_1 = 0) = E(p^2) + E\big(p(1-p)\big) = E(p)$. The comparison yields indeed

$$\mathbb{P}(X_2 = 1|X_1 = 1) - \mathbb{P}(X_2 = 1) = \frac{E(p^2)}{E(p)} - E(p) = \frac{V(p)}{E(p)} > 0.$$

To conclude, reconsidering the Binomial scenarios with and without replacement, observe that dependence occurs in the latter case, while independence occurs in the former case, reversing what occurs for non-random N where $\mathbb{P}(X_2 = 1|X_1 = 1) = \mathbb{P}(X_2 = 1) = N/M$ with replacement, and $\mathbb{P}(X_2 = 1|X_1 = 1) = (N-1)/(M-1) < N/M = \mathbb{P}(X_2 = 1)$ without replacement. This is paradoxical as, otherwise stated, the passage of a fixed number of red balls to a binomially distributed number of red balls reverses the categorization of independence and non-independence.

Acknowledgements

The author wishes to thank and anonymous referee for helpful comments and suggestions.

References

[1] Klarreich, E. (2024), Perplexing the Web, One Probability Puzzle at a Time, in Quanta Magazine.

[2] Otero, D. (2025). A Probability Puzzle, arxiv:2501.03238.

[3] Paterson, M.B. and Stinson, D.R. (2024), Daniel Litt's Probability Puzzle, arxiv:2409.08094.

ÉRIC MARCHAND
Université de Sherbrooke, Département de mathématiques
Sherbrooke Qc, Canada
e-mail: eric.marchand@usherbrooke.ca