Red balls, Green Balls

2024-08-29

This work delves into a probability puzzle involving an urn with an unknown number of red and green balls. Initially, the number of red balls is randomly selected. After observing a red ball on the first draw, the likelihood of drawing another red ball is calculated using Bayesian probability, which adjusts the probability distribution over possible urn configurations. The puzzle is further explored by modifying the setup, introducing a scenario where the number of red balls is determined by repeated coin flips, resulting in a binomial distribution. Theoretical results suggest a $2/3$ probability of drawing a red ball again in the original setup, and a $1/2$ probability when using the binomial model. These outcomes are validated through simulations, highlighting the nuanced effects of initial conditions on probability outcomes and underscoring the importance of rigorous reasoning in probabilistic analysis.

blog: https://tetraquark.vercel.app/posts/redballgreenball/

email: quarktetra@gmail.com

Introduction

Probability is a fascinating yet often perplexing field because it deals with uncertainty and randomness, concepts that can be challenging to intuitively grasp. Many probability problems appear simple on the surface but reveal surprising and counterintuitive results upon closer inspection. For example, the Monty Hall problem, where switching doors in a game show leads to better odds of winning, defies most people's gut instincts. This highlights the importance of rigorous mathematical analysis in probability. Without careful calculation and logical reasoning, our intuitive judgments can lead us astray, underscoring the need for a systematic approach to understanding and applying probability in real-world situations. Today, we will tackle such a problem.

Problem Statement

Quoting from Quanta magazine, [1]

" 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, is it more likely to be red or green (or are the two colors equally likely)? "

— Daniel Litt

Let us define the number of balls in the urn as N. Originally, there are $N+1$ different possible configurations for the urns with the number of red balls from 0 to N. Let is label the urns as u_i, with $i \in \{0, 1, 2, \cdots, N\}$. Since the number of red balls is pulled from a uniform distribution, the probability of getting any u_i is simply $\frac{1}{N+1}$.

First, we will shut-up and apply the rigorous machinery of probability theory. Second, we will do a simulation to double check the results.

Shut up and calculate

Given the first ball is red, the probability that it came from urn u_i (i.e., an urn with i red balls) is

$$P(u_i|x_1 = R) = \frac{P(x_1 = R|u_i)P(u_i)}{P(x_1 = R)}, \tag{1}$$

which is the Bayesian formula. We can compute the denominator as

$$P(x_1 = R) = \sum_{i=0}^{N} P(x_1 = R|u_i)P(u_i) = \sum_{i=0}^{N} \frac{i}{N}\frac{1}{N+1} = \frac{1}{2}. \tag{2}$$

And the numerator as

$$P(x_1 = R|u_i)P(u_i) = \frac{i}{N}\frac{1}{N+1} = \frac{i}{N(N+1)}. \tag{3}$$

Putting Eqs. 2 and 3 back into Eq. 1 gives

$$P(u_i|x_1 = R) = \frac{2i}{N(N+1)}. \tag{4}$$

Note that this is very important. Originally, the probability of getting an urn with i red balls was flat at $\frac{1}{N+1}$. However, the fact that first ball is red implies that it is more likely

that it came from an urn with more red balls, see **?@fig-uniformcase**. Now that we have a probability distribution for having u_i, we can compute the probability of getting a red ball in the second draw:

$$
\begin{aligned}
P(x_2 = R) &= \sum_{i=0}^{N} P(x_2 = R|u_i)P(u_i|x_1 = R) = \sum_{i=0}^{N} \frac{i-1}{N-1}\frac{2i}{N(N+1)} \\
&= \frac{2}{(N-1)N(N+1)} \sum_{i=0}^{N} (i^2 - i) \\
&= \frac{2}{(N-1)N(N+1)} \left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)}{2} \right] \\
&= \frac{2N+1}{3(N-1)} - \frac{1}{N-1} = \frac{2}{3}.
\end{aligned}
\tag{5}
$$

Therefore, getting a red ball is twice as likely as getting a green ball, independent of how many balls there are in the urn.

A tweak

The statement *"All you know is that someone determined the number of red balls by picking a number between zero and 100 from a hat"* is so critical. In order to show why that is, let's revise the preparation of the urn as follows: *"The urn is prepared by adding balls to the urn one by one by flipping a coin. If the coin is heads, you add a red ball, a green one otherwise."* This will completely transform the question. The number of red balls in the urn will have the binomial probability density:

$$
f_b(i, N) = \binom{N}{i} p^i (1-p)^{N-i}.
\tag{6}
$$

Now, we just need to redo the math. Given the first ball is red, the probability that it came from urn u_i (i.e., an urn with i red balls) is

$$
P(u_i|x_1 = R) = \frac{P(x_1 = R|u_i)P(u_i)}{P(x_1 = R)},
\tag{7}
$$

which is still the Bayesian formula. We can compute the denominator as

$$
P(x_1 = R) = \sum_{i=0}^{N} P(x_1 = R|u_i)P(u_i) = \sum_{i=0}^{N} f_b(i, N)\frac{i}{N} = \frac{1}{N}\langle i \rangle = p.
\tag{8}
$$

For a fair coin used in the preparation $p = 1/2$. The numerator reads

$$
P(x_1 = R|u_i)P(u_i) = \frac{i}{N}f_b(i, N).
\tag{9}
$$

Putting Eqs. 8 and 9 back into Eq. 7 gives

$$P(u_i|x_1 = R) = \frac{1}{Np}if_b(i, N).\tag{10}$$

Now that we have a probability distribution for having u_i, we can compute the probability of getting a red ball in the second draw:

$$
\begin{aligned}
P(x_2 = R) &= \sum_{i=0}^{N} P(x_2 = R|u_i)P(u_i|x_1 = R) \\
&= \frac{1}{pN(N-1)} \sum_{i=0}^{N} (i-1)if_b(i, N) = \frac{1}{pN(N-1)} \left(\langle i^2 \rangle - \langle i \rangle\right) \\
&= \frac{1}{pN(N-1)} \left(\langle i \rangle^2 + \sigma^2 - Np\right) = \frac{1}{pN(N-1)} \left(N^2 p^2 + Np(1-p) - Np\right) \\
&= \frac{1}{pN(N-1)} N(N-1)p^2 \\
&= p,\tag{11}
\end{aligned}
$$

which is the probability of the coin that was used to build the urn! Therefore, getting a red ball in the second draw is as probable as getting a green one provided that the coin was unbiased.

Simulation

The code simulates the original problem with 100 balls and repeats it 40000 times.

We can now simply count the cases of second ball being red, and the simulation result is 0.667 ± 0.005 with 95% confidence level which is very close to the 66.7 value we predicted in Eq. 5.

Below is the tweaked problem where the balls are decided with coin flip.

We can now simply count the cases of second ball being red, and the simulation result is 0.494 ± 0.005 with 95% confidence level which is very close to the 0.5 value we predicted in Eq. 11. The color of the first ball isn't really a useful information, it barely moves the distribution. In fact, the shift towards the higher values of red ball count exactly cancels the fact that we threw away a red ball in the first draw. It is a perfect cancellation!

[1]   E. Klarreich, "Perplexing the web, one probability puzzle at a time," *Quanta Magazine*, 2024 [Online]. Available: https://www.quantamagazine.org/perplexing-the-web-one-probability-puzzle-at-a-time-20240829/
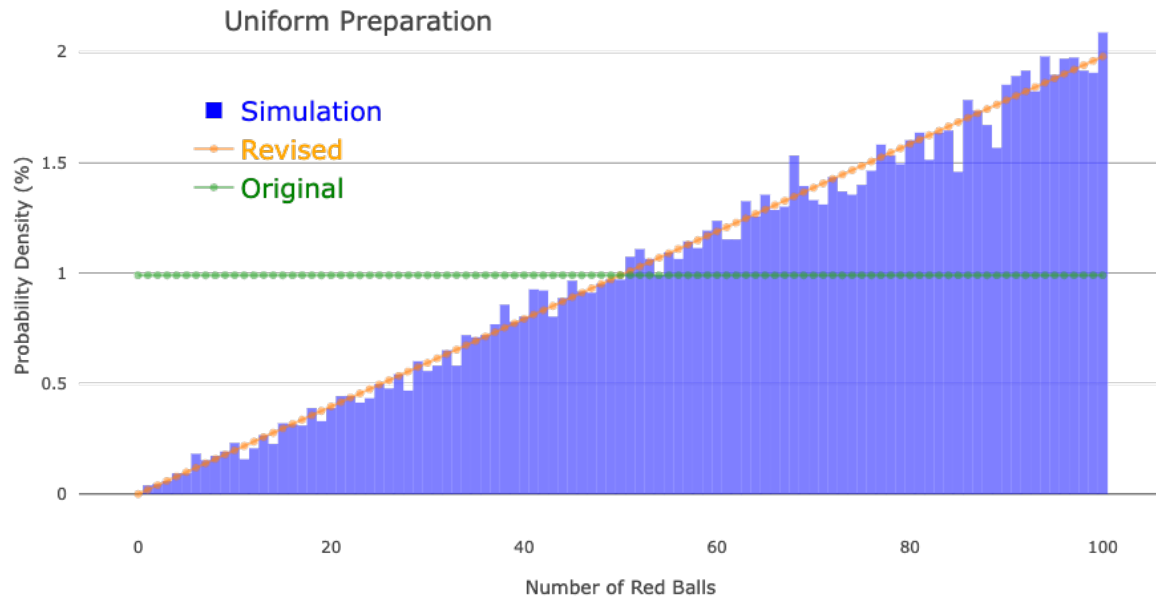
Figure 1: Green: original density, Orange:The probability distribution of having an urn with i red balls given that the first ball was red, Blue: simulation results.
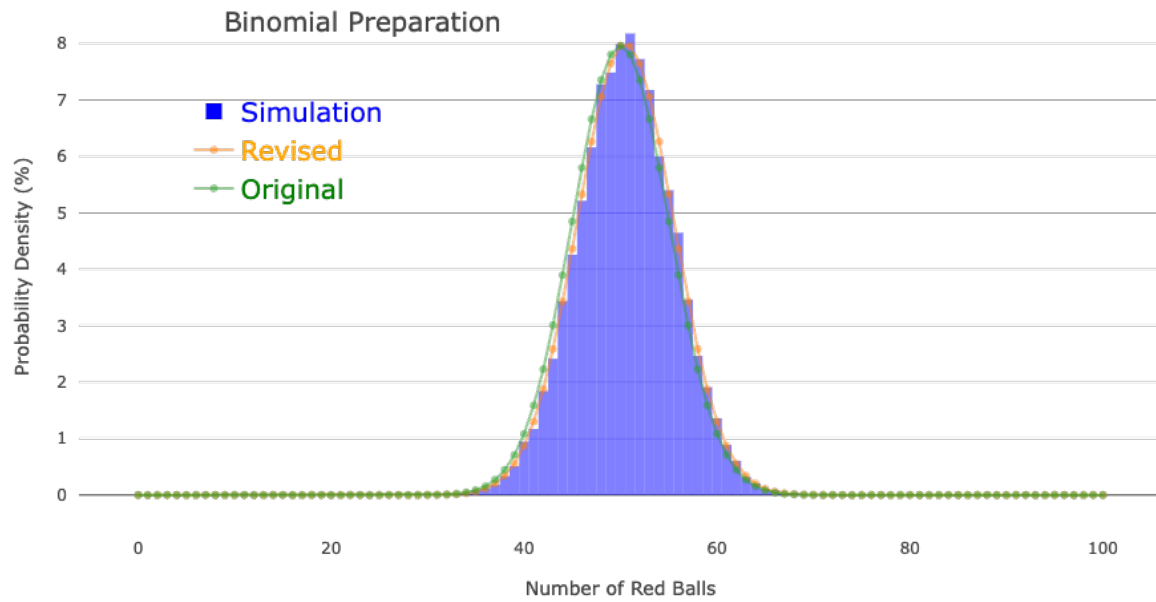

Figure 2: Green: original density, Orange:The probability distribution of having an urn with i red balls inside that the first ball was red, Blue: simulation results.