The Paradox of the First Collision

Gregory Pollard

November 29, 2020

Abstract

In this short report we explain the nature of the above paradox using the context of lottery balls and birthdays, we describe real-world occurrences of this phenomenon and determine the relevant probability formula associated. We'll show this effect in a graphical form and describe how the behaviour of the random variable $X_1^{(n)}$ (the number of repetitions until a singular repeated outcome or "collision" occurs) changes as n increases. Finally, we'll set this result to be a fixed probability and see how k (the number of observed values until a collision occurs) changes as n varies.

1 Introduction

[1] On the 20^{th} of December 1986, an ordinary series of numbers was drawn (15-25-27-30-42-48) in the "Zahlenlotto" in Stuttgart (Germany), exactly the same series of balls was drawn again on the 21^{st} of June 1995 by the same lottery organisation. The "Zahlenlotto" uses balls numbered from 1 to 49, 6 balls are chosen from these randomly and when one is chosen it is not replaced. So the number of ways to choose the 6 balls is:

$$\binom{49}{6} = 13,983,816.$$

Given this many possibilities, and the fact that there had only been roughly 2000 draws overall (balls are drawn once a week), at a glance it seems almost impossible that the same draw could even occur at all within our lifetimes. This report aims to determine a probability formula for how likely this is to occur with different values of n, not just $\binom{49}{6}$; and prove that this is more likely that it might seem. . .

2 Determining the Formula

Let $X_1^{(n)}$ be the random variable that denotes the number of observations needed until a single repeat of an outcome occurs. To explain this in a simpler way, imagine we wanted to find out the probability of at least two people in a room of M people having the same birthday, so $n = 365$. Say we number the days of the year ($1 = 01/01/xx$, $2 = 02/01/xx$, $\dots 365 = 31/12/xx$)[1], and so the birthdays of these M people gives the sequence:

$$72, 55, 12, 178, 42, 12, 230 \dots$$

In this case, $X_1^{(365)} = 5$ since we needed 5 repeats after the first initial observation to achieve the same birthday. Now we need to determine a formula for finding $P[X_1^{(n)} \leq k]$, for some integer k such that $k \leq n$. So we begin by first finding the probability that *no* collision occurs in a general case where there's n possible values, and we observe k of them. That is like asking: "what is the probability that we don't see any collisions after k observations?"[2]. To do this we consider the first observation when $k = 1$:

$$P(a_1) = 1 \qquad \text{Where } a_1 \in \{a_1, a_2, \dots a_k\}.$$

Now consider the probability that the second observation doesn't equate to the first, and also the probability of the third observation not being equal to either of the first two:

$$P(a_1 \neq a_2) = \frac{n-1}{n} = 1 - \frac{1}{n}, \qquad P(a_1 \neq a_2 \neq a_3) = \frac{n-2}{n} = 1 - \frac{2}{n}. \tag{2.1}$$

We can use 2.1 to show that for the nth observation:

$$P(a_1 \neq a_2 \cdots \neq a_k) = \frac{n-(k-1)}{n} = 1 - \frac{k-1}{n}. \tag{2.2}$$

So from 2.2 we can conclude, given that the observations are independent (which they are), the probability of *no* collisions after k observations is as follows:

$$P[X_1^{(n)} > k] = \prod_{i=1}^{k} P(a_i) = \prod_{i=1}^{k-1} (1 - \frac{i}{n}). \tag{2.3}$$

Finally, we can see that the probability of a singular collision after k observations in n values is just the compliment of 2.3:

$$P[X_1^{(n)} \leq k] = 1 - \prod_{i=1}^{k} P(a_i) = 1 - \prod_{i=1}^{k-1} (1 - \frac{i}{n}) \qquad \forall k, n \in \mathbb{N}. \tag{2.4}$$

[1]Excluding $29/02/xx$ (the date added on a leap year).

Note that this is a CMF (Cumulative Mass Function); this is because it takes positive integers as an input, and $P[X_1^{(n)} \leq k] \to 1$ as k increments. Also, as we increase our k, we are checking that it's not equal to every other previously observed value. So if we sub a value for k in, we compare all the observed values with each other up to k, so then at this value we will see a single collision with a previous observation. The CMF in 2.5 shows the full definition:[2]

$$P[X_1^{(n)} \leq k] = \begin{cases} 1 - \prod_{i=1}^{k-1}(1 - \frac{i}{n}) & k \in \mathbb{N} \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

3 A Graphical Representation

Now that we have found a suitable formula in 2.4 for the likelihood of a single collision in a set of n objects, now we can show how the random variable $X_1^{(n)}$ changes behaviour as n increases. Figure 1 shows the probability of a collision as the number of observed values k increases, this is for the birthday problem where $n = 365$. Figure 2 shows the same formula applied to the lottery ball example, where $n = \binom{49}{6}$.

Note how the CMFs have been converted to density functions as it is easier to interpret graphically, we still want to model k as a natural number in reality, however we round the values of k in the set $\{k : k > 0, k \in \mathbb{R}, k \notin \mathbb{N}\}$ to the function value found at the nearest natural number of k. In doing this we have made 2.5 into a step function, and so the graphs are only an approximation of the exact values $P[X_1^{(n)} \leq k]$ can take.

[2] 2.1, 2.2, 2.3 and 2.4 were all accumulated from source [2].

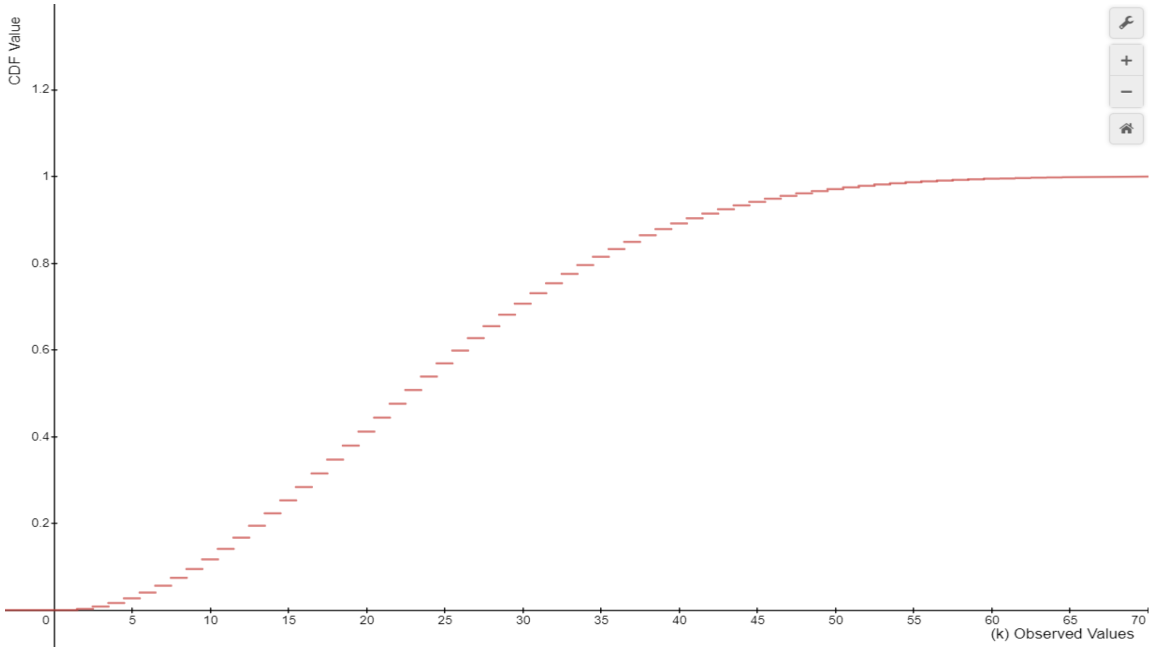

Figure 1: A graph of the probability of a single collision at k with $n = 365$ objects.

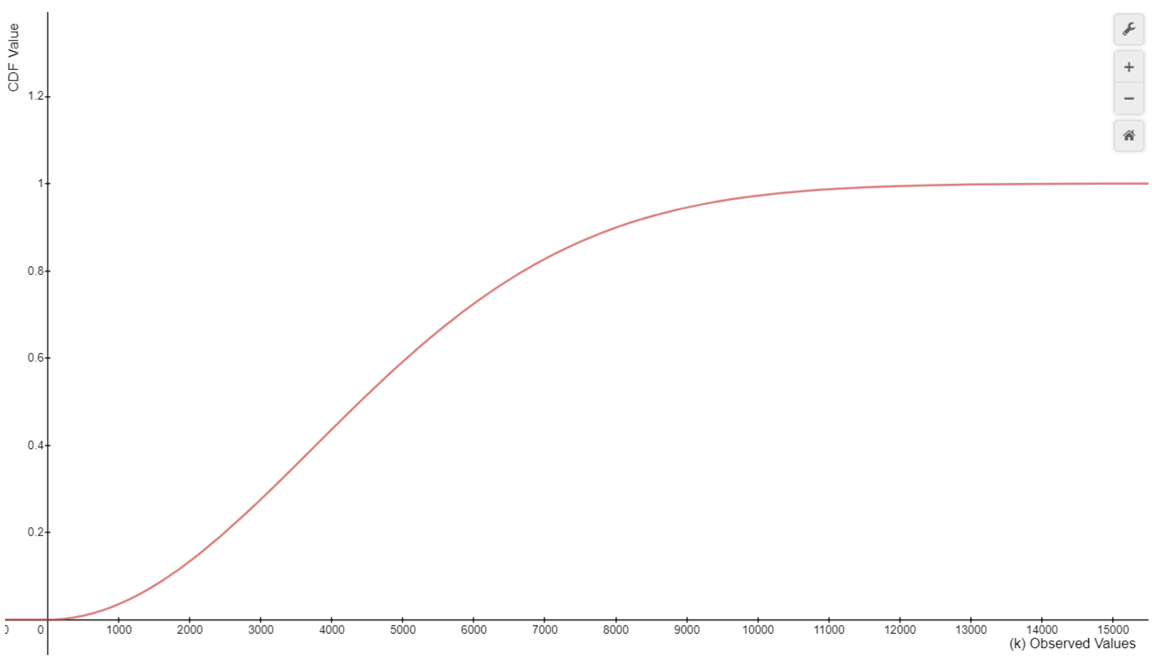

Figure 2: A graph of the probability of a single collision at k with $n = \binom{49}{6}$ objects.

We can see clearly that these CDFs rapidly increase, for example it only takes 23 people in a room for there to be a 50% chance that two of them have the same birthday.

From Figure 1 and Figure 2, we can observe that as n changes, the variable $X_1^{(n)}$ retains its behaviour, however it reaches close to **1** faster compared to the value of n for larger values of n. For example, when $n = 365$ then the probability reaches nearly **1** slowly at around $k = 70$; however for $n = \binom{49}{6}$, it appears to reach **1** almost instantly at around $k = 13,000$ which is very small in proportion to such a large value of n. The chance of a collision starts off small with low values of k, then builds very quickly at a point depending on how large n is; as discussed earlier.

We can now use 2.4 to determine the value of k for a given probability by solving for k; for example in the birthday and lotto scenarios we have:

$$P[X_1^{(365)} \leq k] = 1 - \prod_{i=1}^{k-1}(1 - \frac{i}{365}) = 0.95 \qquad \text{Here, } k = 46$$

$$P[X_1^{\binom{49}{6}} \leq k] = 1 - \prod_{i=1}^{k-1}(1 - \frac{i}{\binom{49}{6}}) = 0.95 \qquad \text{Here, } k = 7,490$$

4 Conclusion

In this investigation we found the formula of the probability that a singular collision will occur after k observed values where each observation can take n different values. This formula is shown formally in 2.5. This was then presented graphically, however it needed to be converted to a density function where $k \in \mathbb{R}$ (as opposed to k being a natural number). This is because plotting a CMF would give function values of infinitely small line width at each k in the domain, so it would have been hard to see properly the behaviour of the variable $X_1^{(n)}$. Using a small value of $n = 365$ and an extremely large value of $n = \binom{49}{6}$, we saw that the CMF increases close to **1** sooner in proportion to the value of n when n is larger and later when n is smaller. The CMF value was then fixed to 0.95 for these values of n to see how k behaves.

The formula derived in 2.5 can now finally explain how the "Zahlenlotto" situation happened. After only 2000 draws (so $k = 2000$), and with $n = \binom{49}{6}$, plugging the numbers produces a 13% chance that the next draw of balls will be the same as one that had already been drawn prior. All of a sudden the chance of this happening seems a lot more likely than one might initially postulate…

References

[1] Elek, G. (2019). Short project: The paradox of the first collision. *file://lancs/homes/38/pollardg/Downloads/GE%20(2).pdf*. 1

[2] redshiftzero(github) (2017). Collision attacks and the birthday paradox. *https://redshiftzero.github.io/birthday-attacks/*. 2, 3