The Cereal Box Problem

Sahand Rabbani

Problem Statement

Consider a brand of cereal that advertises a special prize in each of its cereal boxes. Suppose that there are n possible prizes and that any box of cereal contains each of these n prizes with equal probability, but which prize it contains is unknown until the box is opened. What is the expected number of cereal boxes that one must open to obtain all n prizes?

Solution

We define a random variable X_n that is the number of trials, or cereal box openings, until all n prizes are obtained. We seek $\mathbb{E}[X_n]$. Consider the random variables Y_i for $i = \{1, 2, \ldots, n\}$, where Y_i only has significance after $n - i$ prizes have been obtained. The random variable Y_i is the number of trials required to obtain one of the i remaining prizes. We can thus write X_n as

$$X_n = Y_n + Y_{n-1} + \cdots + Y_1 = \sum_{i=1}^{n} Y_i$$

By the linearity of expectation, we have

$$\mathbb{E}[X_n] = \mathbb{E}[Y_n] + \mathbb{E}[Y_{n-1}] + \cdots + \mathbb{E}[Y_1] = \sum_{i=1}^{n} \mathbb{E}[Y_i]$$

We note that each Y_i is a geometric random variable where the probability of success p_i is $\frac{i}{n}$. The simplest case is $Y_n = 1$. When none of the prizes have been obtained, the expected number of trials to obtain one of the n prizes is deterministic: open one cereal box and you have one of these n prizes. In general, the expected value of a geometric random variable with probability of success p is $\frac{1}{p}$. Thus, the expected value of X_n is

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} \mathbb{E}[Y_i] = \sum_{i=1}^{n} \frac{1}{i/n} = n \sum_{i=1}^{n} \frac{1}{i} = nH_n$$

where H_n is the nth harmonic number defined as the sum of the reciprocals of the first n natural numbers.