5. Coupon Collector's Problem

There are many cereal boxes, and in each cereal box with equal probability there is one of the n coupons $C_1, C_2, C_3, \ldots C_n$.

How many boxes do we have to buy to collect all n coupons?

X_i - Random variable that is defined to be equal to the number of purchases needed to get i^{th} coupon.
Our answer is $E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$

$X_1 = 1$ as the first box will always result in a new coupon for us

Finding X_i

We know the probability of success when you're opening a box for getting the second new coupon is

$$Pr[\text{success}] = \frac{n-1}{n}$$

Therefore the expected number of trials for getting the second coupon is

$$E[X_2] = \frac{n}{n-1}$$

Hence, for general X_i,

$$E[X_i] = \frac{n}{n-(i-1)}$$

Hence using linearity of expectation

$$\sum_{i=1}^{n} E[X_i] = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} = n \ln n$$

Hence the expected number of boxes needed is

$$n \ln n$$

Probability of deviation from expected value

We've seen Markov's inequality, that

$$Pr[X \geq a] \leq \frac{E[X]}{a}$$

For Coupon collector's problem, if we wanted to find the probability that it takes $\geq 2 \cdot E[X]$ coupons,

$$P[X \geq 2E[X]] \leq \frac{E[X]}{2E[X]} = \frac{1}{2}$$

However, that may not be good enough every time, and it isn't in this case at least. Hence, we have to use Chebyshev's Inequality, which says -

$$P(|x - \mu| \geq a) \leq \frac{Var(X)}{a^2}$$

Finding tighter bound for Coupon Collector problem

Variance of geometric distribution $= \frac{1-p}{p^2}$
For independent random variables $X_1, X_2, \ldots, X_n$, we can sum the variances

$$\begin{aligned}
Var[X] &= \sum_{i=1}^{n} Var[X_i] \\
&= \sum_{i=1}^{n} \frac{(1-p_i)}{p_i^2} \\
&= \sum_{i=1}^{n} \frac{\left(1 - \left(\frac{n-i+1}{n}\right)\right)}{\left(\frac{n-i+1}{n}\right)^2} \\
&= \sum_{i=1}^{n} \frac{\left(\frac{i-1}{n}\right)}{\frac{(n-i+1)^2}{n^2}} \\
&= \sum_{i=1}^{n} \frac{n(i-1)}{(n-i+1)^2} \\
&= \sum_{j=1}^{n} \frac{n(n-j)}{j^2} & (j = n-i+1) \\
&= \left(n^2 \sum_{j=1}^{n} \frac{1}{j^2}\right) - \left(n \sum_{j=1}^{n} \frac{1}{j}\right) \\
&\leq \frac{n^2\pi^2}{6} - n \cdot H_n \\
&\leq \frac{n^2\pi^2}{6}
\end{aligned}$$

Now, finding the probability that it takes $\geq 2E[X] \approx 2n \lg n$ coupons is

$$Pr[X \geq 2E[X]] \leq Pr[|X - E[X]| \geq E[X]]$$

Using Chebyshev's Inequality, $a = E[X]$.

$$\begin{aligned} Pr[|X - E[X]| \geq E[X]] &\leq \frac{Var(X)}{E[X]^2} \leq \frac{\pi^2 n^2}{6n^2 H_n^2} \\ &= \frac{\pi^2}{6H_n^2} \leq \frac{2}{\ln^2(n)} \end{aligned}$$