29 Jan 2025

The Coupon Collector Problem

Def The geometric distribution with parameter p is the distribution of the # of coin tosses until (and including) the first time heads is tossed, when tossing a coin with $\Pr(\text{heads}) = p$, and independent coin-toss outcomes.

$$\Pr(X=k) = (1-p)^{k-1} \cdot p$$

k-1 tosses yield tails... ...and k^{th} toss is heads.

Expectation of geometric rand var

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \cdot \Pr(X=k)$$

$$= \sum_{k=1}^{\infty} k\,(1-p)^{k-1} \cdot p$$

Let $q = 1-p$

$$= p \sum_{k=1}^{\infty} k\, q^{k-1}$$

$$= p \frac{d}{dx}\left[\sum_{k=1}^{\infty} x^k\right]_{x=q} = p \frac{d}{dx}\left[\frac{x}{1-x}\right]_q$$

$$= p \frac{d}{dx}\left[1 + \frac{x}{1-x}\right]_{x=q}$$

$$= p \frac{d}{dx}\left[\frac{1}{1-x}\right]_{x=q}$$

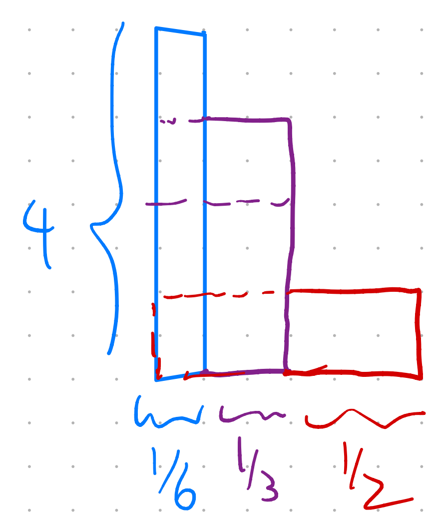
$$= \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}.$$

Fact: If X is a random variable taking non-negative integer values,

$$\boxed{E[X] = \sum_{j=0}^{\infty} \Pr(X > j)}$$

Proof by picture. Suppose $X = \begin{cases} 1 & \text{w. prob. } 1/2 \\ 3 & \text{w. prob } 1/3 \\ 4 & \text{w. prob } 1/6 \end{cases}$

$$E[X] = \sum k \cdot \Pr(X = k)$$



For geometric RV,

$$\Pr(X > j) = \Pr(\text{first } j \text{ coin tosses are tails})$$
$$= (1-p)^j$$

$$E[X] = \sum_{j=0}^{\infty} \Pr(X > j) = \sum_{j=0}^{\infty} (1-p)^j = \frac{1}{1-(1-p)} = \frac{1}{p}$$

Last derivation. (principle of deferred decisions)

If first coin toss is H, $X = 1$.

If — — — T, $X = 1 + Y$

where Y is $\mathrm{Geom}(p)$.

$$\mathbb{E}[X] = \mathbb{E}[X \mid \text{first toss } H] \cdot \Pr(H)$$
$$+ \mathbb{E}[X \mid \text{first toss } T] \cdot \Pr(T)$$

$$= 1 \cdot p + \left(1 + \mathbb{E}(X)\right) \cdot (1-p)$$

$$p \cdot \mathbb{E}(X) = 1 \cdot p + 1 \cdot (1-p) = 1$$

$$\mathbb{E}(X) = \frac{1}{p}.$$

Variance of geometric RV.

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \left(\mathbb{E}(X)\right)^2 \quad = \frac{1}{p^2}$$

$$\mathbb{E}[X^2] = \mathbb{E}[X^2 \mid \text{first toss is } H] \cdot p$$
$$+ \mathbb{E}[X^2 \mid \text{first toss is } T] \cdot (1-p)$$

$$= 1 \cdot p + \mathbb{E}[(1+Y)^2] \cdot (1-p) \qquad Y \sim \mathrm{Geom}(p)$$

$$= 1 \cdot p + \mathbb{E}[(1+X)^2] \cdot (1-p)$$

$$= p + \Big(\underbrace{\mathbb{E}[1]}_{=1} + \underbrace{\mathbb{E}[2X]}_{=\frac{2}{p}} + \mathbb{E}[X^2]\Big)(1-p)$$

$$\mathbb{E}[X^2] = p + 1-p + \frac{2(1-p)}{p} + (1-p)\,\mathbb{E}[X^2]$$

$$p \cdot \mathbb{E}[X^2] = 1 + \frac{2-2p}{p}$$

$$\mathbb{E}(X^2) = \frac{1}{p} + \frac{2}{p^2} - \frac{2}{p} = \frac{2-p}{p^2}$$

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \frac{1-p}{p^2}$$

COUPON COLLECTOR!

Phase 1 ends at τ_1 = first time at least 1 bin is occupied $= 1$

⋮

Phase k ends at τ_k = first time at least k bins are occupied.

If $\tau_k \leq t < \tau_{k+1}$ it means

exactly k bins are occupied at time t

$\Longrightarrow$ Pr(ball thrown at $t+1$ occupies new bin)

$$= \frac{n-k}{n}$$

The rand var $X_{k+1} = \tau_{k+1} - \tau_k$ is geom. distr. with parameter $p = \frac{n-k}{n}$, independent of $X_1, X_2, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n$.

$$E[\tau_n] = E\left[\tau_1 + (\tau_2 - \tau_1) + (\tau_3 - \tau_2) + \cdots + (\tau_n - \tau_{n-1})\right]$$

$$= E(X_1 + X_2 + \cdots + X_n)$$

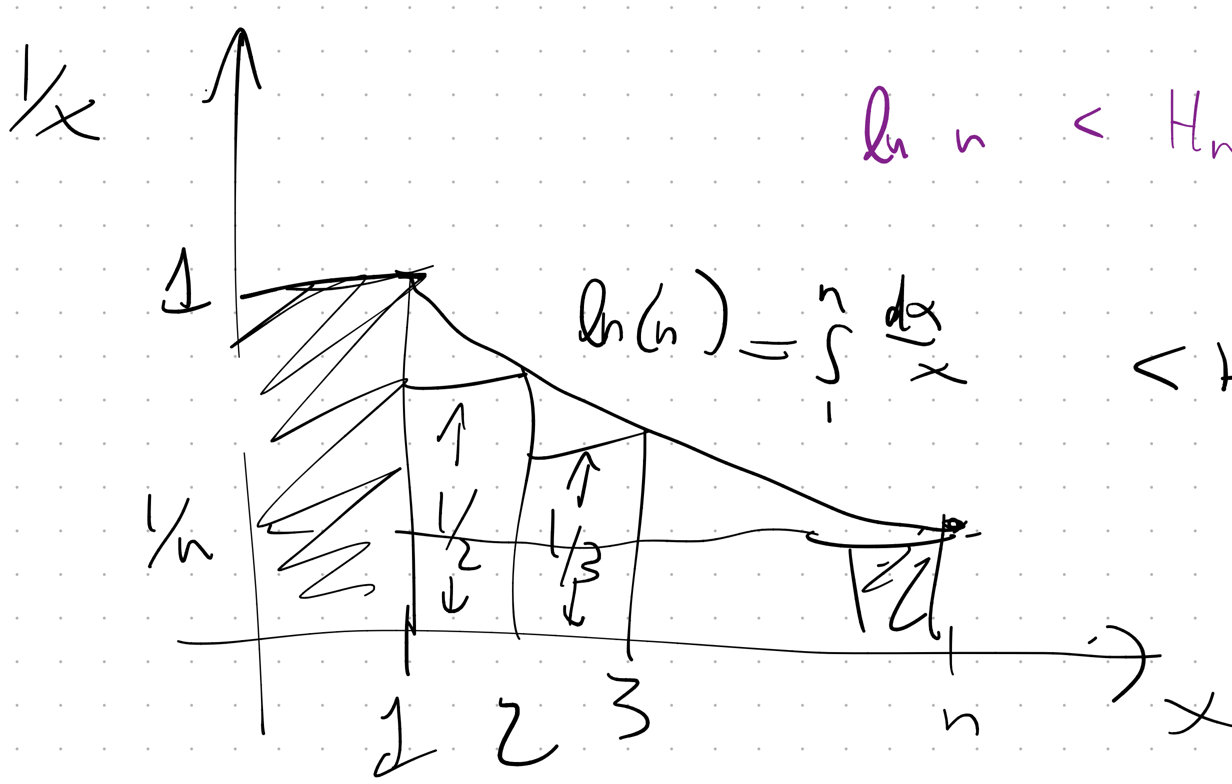
$$= \sum_{k=1}^{n} E\left(\mathrm{Geom}\left(\frac{n-(k-1)}{n}\right)\right)$$

$$= \sum_{k=1}^{n} \frac{n}{n-(k-1)} = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1}$$

$$= n \cdot \left[1 + \frac{1}{2} + \cdots + \frac{1}{n}\right]$$

$$= n \cdot H_n$$ ← n^{th} harmonic number.

$$\ln n < H_n < 1 + \ln n.$$



$$\ln(n) = \int_1^n \frac{dx}{x} < H_n < 1 + \int_1^n \frac{dx}{x} = 1 + \ln(n)$$

Suppose $1/p \in \mathbb{N}$.

$$\Pr\left(\text{Geom}(p) > \frac{1}{p}\right) = (1-p)^{1/p}$$

$$< \left(e^{-p}\right)^{1/p} = e^{-1} \approx 0.36\ldots$$

Let $m_{\text{coupon}}(n)$ be the smallest m s.t.

$\Pr(\tau_n > m) \leq \frac{1}{2}$.

Estimating $m_{\text{coupon}}(n)$ ......

To show $\Pr(\tau_n$ far from $\mathbb{E}\tau_n$ is small)

we use Chebyshev's Ineq.

$$\text{Var}(\tau_n) = \text{Var}(X_1 + X_2 + \cdots + X_n)$$

$$= \text{Var}(X_1) + \cdots + \text{Var}(X_n)$$

$$= \sum_{k=1}^{n} \frac{1 - \left(\frac{n-k+1}{n}\right)}{\left(\frac{n-k+1}{n}\right)^2}$$

$$< \sum_{k=1}^{n} \frac{1}{\left(\frac{n-k+1}{n}\right)^2} = n^2 \cdot \underbrace{\left[1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2}\right]}_{< 2}$$

$$< 2n^2$$

$\forall x$

$$\Pr\left(\left|\tau_n - \mathbb{E}\tau_n\right| > x\right)$$

$$= \Pr\left(\left(\tau_n - \mathbb{E}\tau_n\right)^2 > x^2\right)$$

$$< \frac{\mathbb{E}\left(\left(\tau_n - \mathbb{E}\tau_n\right)^2\right)}{x^2} < \frac{2n^2}{x^2}.$$

$$\Pr\left(\left|\tau_n - \mathbb{E}\tau_n\right| > 2n\right) < \frac{2n^2}{4n^2} = \frac{1}{2}.$$

The event $\tau_n > \overset{\mathbb{E}\tau_n}{n \cdot H_n} + 2n$

cannot have probability $> \frac{1}{2}$.

$$\therefore \quad m_{\text{coupon}}(n) < n H_n + 2n$$

$$< n \ln(n) + 3n$$

Also

$$m_{\text{coupon}}(n) > n \ln(n) - 2n.$$