

2. Moments and Deviations

In our first lecture we saw that the expected running time of RandQS can be bound very well.

Question: How close is the running time to the expectation?

We shall illustrate this with some examples.

2.1 Coupon collector's problem.

A brand of Cereal is running a promotion by shipping each of its boxes with a coupon from a collection of n coupons. A shopper who collects all of those n coupons wins a prize.

Let us suppose that the brand places coupons in the box uniformly at random, from the n available coupons.

Question: How many boxes must a shopper buy to collect all the coupons?

Coupons	c_1	c_2	c_3	c_4	\dots	c_n
---------	-------	-------	-------	-------	---------	-------

Time it is picked	t_1	t_2	t_3	t_4	\dots	t_n
----------------------	-------	-------	-------	-------	---------	-------

What is an estimate on $\max_i \{t_i\}$?

Let X be the random variable defined to be equal to the no. of boxes to be bought to collect each of the n distinct coupons. Let $\hat{c}_1, \dots, \hat{c}_X$ be the coupons drawn in the X purchases. (We shall refer to each purchase as a trial.)

In other words $\{\hat{c}_1, \dots, \hat{c}_X\}$ contains all of $\{c_1, \dots, c_n\}$.

Let X_i be the random variable defined to be equal to the number of trials needed to collect the i^{th} coupon, after $(i-1)^{\text{th}}$ coupon was collected.



From the above statement, we get that

$$X = \sum_{i=1}^n X_i.$$

From linearity of expectation,

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i]$$

Now let us estimate what $\mathbb{E}[X_i]$ is for some arbitrary i .

Claim: $\mathbb{E}[X_i] = \frac{n}{n-i+1}$

Proof: Probability of picking a new coupon in each of the trials before picking the i^{th} coupon is equal to $p_i = \frac{n-i+1}{n}$. Note that the random variable X_i is "geometrically distributed". Thus, $\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$.

Digression 1 (Bernoulli random variables and Geometric distributions)

- Suppose we run an experiment that succeeds with a prob of p and fails with a prob of $1-p$. Let Y be a random variable s.t $Y=1$ if the experiment succeeds and it is

0 otherwise. Then Y is called a Bernoulli random variable (also called Indicator r.v.). Further,

$$\mathbb{E}[Y] = \Pr[Y=0] = p.$$

- Suppose X is a random variable defined to be equal to the no. of trial needed to succeed. Thus,

$$\Pr[X=i] = (1-p)^{i-1} \cdot p \text{ and } \Pr[X \geq i] = \sum_{j \geq i}^{\infty} (1-p)^{j-1} \cdot p = (1-p)^{i-1}$$

Further $\mathbb{E}[X] = \sum_{i=1}^{\infty} \Pr[X \geq i]$ Also, $\text{Var}[X] = \frac{1-p}{p^2}$.

$$= \sum_{i=1}^{\infty} (1-p)^{i-1}$$

(Can be proved similarly)

$$= \underbrace{\frac{1}{1-(1-p)}}_{\text{Further, } \forall k, l \in \mathbb{N}} = \frac{1}{p}.$$

$\Pr[X=k+l \mid X > l] = \Pr[X=k]$

↑
(memoryless)

Back to our coupon collector's problem.

$$\begin{aligned} \mathbb{E}[X] &= \sum_{i=1}^n \mathbb{E}[X_i] \\ &= \sum_{i=1}^n \frac{n}{n-i+1} \\ &= n \sum_{j=1}^n \frac{1}{j} = n H_n. \end{aligned}$$

Recall that $H_n \approx \ln n + \Theta(1)$.

$\downarrow j=n-i+1$

Thus, in expectation, a shopper needs to purchase roughly $n \ln n + \Theta(n)$ many boxes of cereal.

Remark: We still have not answered what the behaviour of X is with respect to $\mathbb{E}[X]$.

Digression 2 (Markov Inequality)

Let Y be a r.v assuming only non-negative values.

Then for all $t \in \mathbb{R}_{\geq 0}$,

$$\Pr[Y \geq t] \leq \frac{\mathbb{E}[Y]}{t}$$

$$\begin{aligned} \text{Follows from } \mathbb{E}[Y] &= \sum_{i=0}^{\infty} i \cdot \Pr[Y=i] \geq \sum_{i \geq t} i \cdot \Pr[Y=i] \\ &\geq t \cdot \sum_{i \geq t} \Pr[Y=i] \\ &= t \cdot \Pr[Y \geq t]. \end{aligned}$$

By applying Markov Inequality, we get that

$$\begin{aligned} \Pr[X \geq 2nH_n] &\leq \frac{\mathbb{E}[X]}{2nH_n} = \frac{nH_n + O(n)}{2nH_n} \\ &= \frac{1}{2} + o(1). \end{aligned}$$

Question: Can we make this "tail probability" more "tighter"?

Digression 3 (Chebyshev's inequality)

Let X be a r.v with expectation μ_x and standard deviation σ_x . Then for any $t \in \mathbb{R}_{\geq 0}$

$$\Pr[|X - \mu_x| \geq t\sigma_x] \leq \frac{1}{t^2}$$

$$\sigma_x = \sqrt{\text{Var}[X]}.$$

Can be proved using the fact that $\Pr[|X - \mu_x| \geq t\sigma_x] \geq \Pr[|X - \mu_x|^2 \geq t^2\sigma_x^2]$ and applying Markov on $Y = (X - \mu_x)^2$. $\mathbb{E}[Y] = \sigma_x^2$.

Recall that $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Further, for independent r.v.s X_1, \dots, X_n ,

$$\text{Var}[X] = \sum_{i=1}^n \text{Var}[X_i]$$

$$= \sum_{i=1}^n (1 - P_i) \quad \text{From geometric distribution}$$

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

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

$$= \sum_{i=1}^n \frac{n(i-1)}{(n-i+1)^2} \quad j = n-i+1$$

$$= \sum_{j=1}^n \frac{n(n-j)}{j^2}$$

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

$$\leq n^2 \cdot \frac{\pi^2}{6} - n \cdot H_n$$

$$\text{Thus, } \Pr[|X - \mathbb{E}[X]| > t \cdot \underline{o}(n)] \leq \frac{1}{t^2}.$$

Obtained from

$$\sqrt{\text{Var}[X]}$$

In prev analysis, we had $t \sim \Theta(\log n)$. and got a bound of $\frac{1}{2} + o(1)$ but here we get a stronger bound of $\lesssim \frac{1}{\log^2 n}$.

For $k \in \mathbb{N}$, k^{th} moment $m_x^{(k)}$ and k^{th} central-moment $\mu_x^{(k)}$ of a random variable X are defined as follows:

$$m_x^{(k)} = \mathbb{E}[X^k]$$

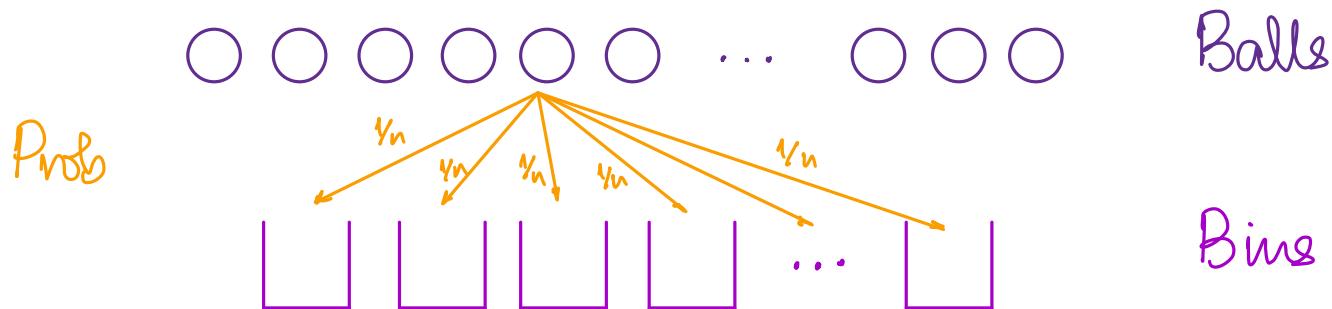
$$\mu_x^{(k)} = \mathbb{E}[(X - \mathbb{E}[X])^k]$$

Expectation is $m_x^{(1)}$ and Variance is $\mu_x^{(2)}$.

2.2 Balls and Bins.

We are given m "indistinguishable" balls and n bins. Each ball is placed in a bin that is chosen uniformly at random.

Question: How are these balls distributed amongst the bins.



Question 1: What is the expected no. of balls in each bin?