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?

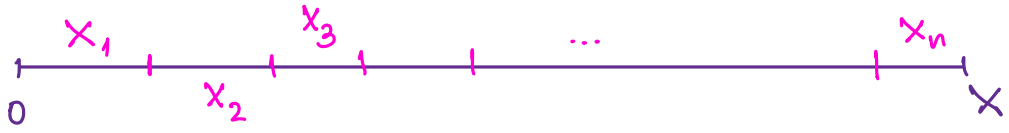
Coupons	C_1	C_2	C_3	C_4	$\cdots$	C_n
Time it is picked	t_1	t_2	t_3	t_4	$\cdots$	t_n

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 $\hat{C}_1, \ldots, \hat{C}_X$ be the coupons drawn in the X purchases. (We shall refer to each purchase as a trial.)

In other words $\{\hat{C}_1, \ldots, \hat{C}_X\}$ contains all of $\{C_1, \ldots, C_n\}$.

Let X_i be the random variable defined to be equal to the number of trials needed to collect the i^{th} coupon, after $(i-1)^{th}$ 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: $\mathbb{E}[X_i] = \frac{n}{n-i}$.

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

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,

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

Further
$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \Pr[X \geq i]$$
$$= \sum_{i=1}^{\infty} (1-p)^{i-1}$$
$$= \frac{1}{1-(1-p)} = \frac{1}{p}.$$

Also, $\mathrm{Var}[X] = \frac{1-p}{p^2}$.

(Can be proved similarly)

Further, $\forall\, k, l \in \mathbb{N}$
$\Pr[X = k+l \mid X > l] = \Pr[X = k]$
↑ (memoryless)

Back to our coupon collector's problem.

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

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

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

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

Follows from $\mathbb{E}[Y] = \sum_{i=0}^{\infty} i \cdot \Pr[Y=i] \geq \sum_{i \geq t} i \cdot \Pr[Y=i]$

$$\geq t \cdot \sum_{i \geq t} \Pr[Y=i]$$

$$= t \cdot \Pr[Y=i].$$

By applying Markov Inequality, we get that

$$\Pr[X \geq 2n H_n] \leq \frac{\mathbb{E}[X]}{2n H_n} = \frac{n H_n + \Theta(n)}{2n H_n}$$

$$= \frac{1}{2} + o(1).$$

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 $t \in \mathbb{R}_{\geq 0}$

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

$\sigma_X = \sqrt{\mathrm{Var}[X]}$.

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 $\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Further, for independent r.vs $X_1, \ldots, X_n$,

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

$$= \sum_{i=1}^{n} \frac{(1-p_i)}{p_i^2}$$ ← From geometric distribution

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

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

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

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

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

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

Thus, $\Pr\left[|X - \mathbb{E}[X]| > t \cdot \underbrace{O(n)}\right] \leq \frac{1}{t^2}$.

Obtained from $\sqrt{\mathrm{Var}[X]}$

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:

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

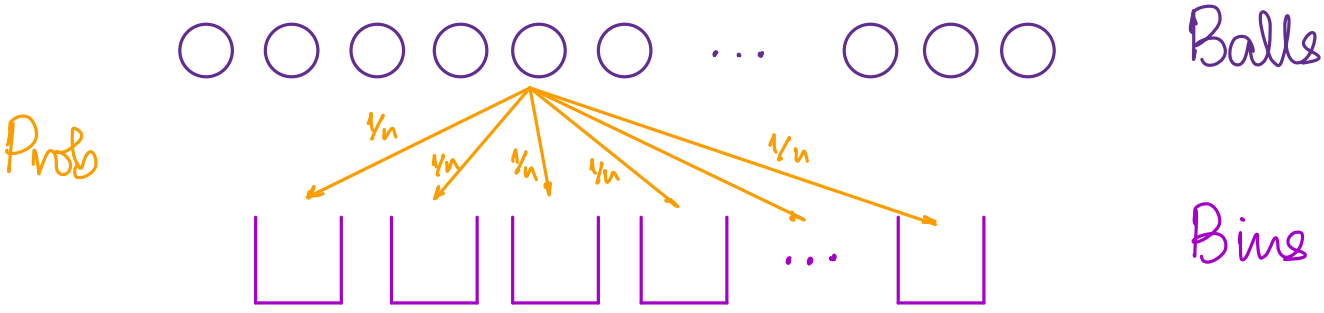
$$\mu_x^{(k)} = \mathbb{E}[(X - \mathbb{E}[X])^k]$$

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?