CS 70 Discrete Mathematics and Probability Theory
Spring 2022 Satish Rao and Koushik Sen **DIS 12B**

1 Student Life

In an attempt to avoid having to do laundry often, Marcus comes up with a system. Every night, he designates one of his shirts as his dirtiest shirt. In the morning, he randomly picks one of his shirts to wear. If he picked the dirtiest one, he puts it in a dirty pile at the end of the day (a shirt in the dirty pile is not used again until it is cleaned). When Marcus puts his last shirt into the dirty pile, he finally does his laundry, and again designates one of his shirts as his dirtiest shirt (laundry isn't perfect) before going to bed. This process then repeats.

(a) If Marcus has n shirts, what is the expected number of days that transpire between laundry events? Your answer should be a function of n involving no summations.

(b) Say he gets even lazier, and instead of organizing his shirts in his dresser every night, he throws his shirts randomly onto one of n different locations in his room (one shirt per location), designates one of his shirts as his dirtiest shirt, and one location as the dirtiest location. In the morning, if he happens to pick the dirtiest shirt, *and* the dirtiest shirt was in the dirtiest location, then he puts the shirt into the dirty pile at the end of the day and does not throw any future shirts into that location and also does not consider it as a candidate for future dirtiest locations (it is too dirty). What is the expected number of days that transpire between laundry events now? Again, your answer should be a function of n involving no summations.

Solution:

(a) The number of days that it takes for him to throw a shirt into the dirty pile can be represented as a geometric RV. For the first shirt, this is the geometric RV with $p = 1/n$. We can see this by noticing that every day the probability of getting the dirtiest shirt remains $1/n$.

We'll call X_i the number of days that go until he throws the ith shirt into the dirty pile. Since on the ith shirt, there are $n-i+1$ shirts left, we get that $X_i \sim \text{Geometric}(1/(n-i+1))$. The number of days until he does his laundry is a sum of these variables. Therefore, we can get the following result:

$$\mathbb{E}[X] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n}(n-i+1) = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

(b) For this part we can use a similar approach but the probability for X_i becomes $1/(n-i+1)^2$. This is because the dirtiest shirt falls into the dirtiest spot with probability $1/(n-i+1)$ and we pick it after that with probability $1/(n-i+1)$, so the probability of picking the dirtiest shirt from the dirtiest spot for the ith shirt is $1/(n-i+1)^2$. Using the same approach, we get the following sum:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} (n-i+1)^2 = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

2 Planetary Party

(a) Suppose we are at party on a planet where every year is 2849 days. If 30 people attend this party, what is the exact probability that two people will share the same birthday? You may leave your answer as an unevaluated expression.

(b) From lecture, we know that given n bins and m balls, $\mathbb{P}\big[\text{no collision}\big] \approx \exp\big(-m^2/(2n)\big)$. Using this, give an approximation for the probability in part (a).

(c) What is the minimum number of people that need to attend this party to ensure that the probability that any two people share a birthday is at least 0.5? You can use the approximation you used in the previous part.

(d) Now suppose that 70 people attend this party. What the is probability that none of these 70 individuals have the same birthday? You can use the approximation you used in the previous parts.

Solution:

(a) Let's compute the probability that no two partygoers have the same birthday. We know the second person at the party cannot share a birthday with the first person, the third person at the party cannot share a birthday with the first two, etc. Thus

$$\mathbb{P}\big[\text{no collision}\big] = \left(1 - \frac{1}{2849}\right)\left(1 - \frac{2}{2849}\right)\left(1 - \frac{3}{2849}\right)\cdots\left(1 - \frac{29}{2849}\right)$$

Thus $\mathbb{P}[\text{collision}] = 1 - \mathbb{P}[\text{no collision}] = 1 - \left(1 - \frac{1}{2849}\right)\left(1 - \frac{2}{2849}\right)\left(1 - \frac{3}{2849}\right)\cdots\left(1 - \frac{29}{2849}\right)$.

(b) From lecture, we know that given n bins and m balls, $\mathbb{P}\big[\text{no collision}\big] \approx \exp\big(-m^2/(2n)\big)$. Therefore in this case, if we want to find the probability of collision, we must find $1 - \mathbb{P}\big[\text{no collision}\big]$.

$$\mathbb{P}\big[\text{no collision}\big] = \exp\left(-\frac{30^2}{2 \cdot 2849}\right) = 0.854$$

This means that there is a 0.146 chance that two people share the same birthday in the group of 30.

(c) Rephrasing the question in terms of balls and bins, we want to find the minimum number of balls (m) such that there is at least 0.5 probability of collision when we have $n = 2849$ bins, which is the same as at **most** 0.5 probability of **no** collisions.

$$\begin{aligned}\mathbb{P}[\text{no collisions}] \approx \exp\left(\frac{-m^2}{2n}\right) &\le 0.5 \\ \implies \frac{-m^2}{2n} &\le \ln 0.5 \\ \implies m &\ge \sqrt{(-2\ln 0.5)n} \\ &= 62.845\end{aligned}$$

Since m must be an integer which is at least 62.485, we need at least $\boxed{63}$ people at the party.

(d) Once again we need to find $\mathbb{P}\left[\text{no collisions}\right]$ given that $m = 70$.

$$\mathbb{P}\left[\text{no collision}\right] = \exp\left(-\frac{70^2}{2 \cdot 2849}\right) = 0.423$$

There is about a 42% chance that 70 people don't share the same birthday.

3 Throwing Balls into a Depth-Limited Bin

Say you want to throw n balls into n bins with depth $k-1$ (they can fit $k-1$ balls, after that the bins overflow). Suppose that n is a large number and $k = 0.1n$. You throw the balls randomly into the bins, but you would like it if they don't overflow. You feel that you might expect not too many balls to land in each bin, but you're not sure, so you decide to investigate the probability of a bin overflowing.

(a) Count the number of ways we can select k balls to put in the first bin, and then throw the remaining balls randomly. You should assume that the balls are distinguishable.

(b) Argue that your answer in (a) is an upper bound for the number of ways that the first bin can overflow.

(c) Calculate an upper bound on the probability that the first bin will overflow.

(d) Upper bound the probability that some bin will overflow. [*Hint*: Use the union bound.]

(e) How does the above probability scale as n gets really large?

Solution:

(a) We choose k of the balls to throw in the first bin and then throw the remaining $n-k$, giving us $\binom{n}{k}n^{n-k}$.

(b) Certainly any outcome of the ball-throwing that overflows the first bin is accounted for – we can simply choose the first k balls that land in the first bin and then simulate the rest of the outcome via random throwing. However, we are potentially overcounting: if $k+1$ balls go in the first bin, we have many choices for which k of them that could have been the "chosen" ones, and we count each one of these choices as distinct. However, they correspond to the same configuration, namely the one where $k+1$ balls are in the first bin. Hence we get an upper bound.

(c) We divide by the total number of ways the balls could have fallen into the bins, with order, so we get

$$\frac{\binom{n}{k} n^{n-k}}{n^n} = \frac{\binom{n}{k}}{n^k}.$$

(d) Let A_i denote the event that bin i overflows. By symmetry $\mathbb{P}[A_i] = \mathbb{P}[A_1]$ for all i. By the union bound we have

$$\mathbb{P}[\cup_i A_i] \le \sum_{i=1}^{n} \mathbb{P}[A_i] \le n\mathbb{P}[A_i] \le n \cdot \frac{\binom{n}{k}}{n^k}.$$

(e) We get

$$n \cdot \frac{\binom{n}{k}}{n^k} = n \cdot \frac{n \cdot (n-1) \cdots (n-k+1)}{k! n^k} \le n \cdot \frac{n^k}{k! n^k} = \frac{n}{k!} = \frac{n}{(0.1n) \cdot (k-1)!} = \frac{10}{(0.1n-1)!}.$$

Clearly, as n gets large this probability is going to 0. Note that this same analysis would work with $k = cn$ for any constant $0 < c < 1$. Hence, using some very coarse upper bounds, we can see that as the number of balls and bins grows, we have that it is very unlikely that we get a constant fraction of the balls in any single bin.