CS6550 : Scaling to Big Data : Homework

24 May, 2018

These are sample problems for the end semester exam, as well as Homework. You may turn in the solutions to these problems by June 6, 2018. Feel free to ask me for hints or if you need more time. Good Luck!

- Markov's inequality: For any non negative r.v. X, and $\alpha > 0$, $\Pr[X \geq \alpha] \leq E(X)/\alpha$.
- Chebysehv's inequality: For any r.v. Y, $\Pr[|Y - E(Y)| \geq \alpha] \leq \text{Var}(Y)/\alpha^2$.
- Chernoff's bound: Let $X_1, X_2, \ldots, X_t$, be i.i.d random var. from $\{0, 1\}$ with expectation μ. If $X = \left(\sum X_i\right)/t$ and $0 < \delta < 1$, then
$$\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\mu t\delta^2/3}.$$

1. Isaac is taking a 5 question multiple choice test in which each question has 4 answers including just one correct answer. He needs to answer at least 4 of the questions correctly in order to pass the test. What is the probability that he will pass?

2. Let A and B be two events such that $\Pr(A) = 0.6$ and $\Pr(B) = 0.8$. Can these two events be mutually exclusive? Answer Yes/No with justifications.

3. A company sells term life insurance policies. The premium for insuring a the life of 30 year old male for Rs. 1,00,00,000 (one crore) for one year is Rs. 15000. The probability that a 30 year old male survives for one year is 0.9986. What is the expected profit for the company for one such policy?

4. Suppose A and B are pairwise independent events, and suppose B and C are pairwise independent events. Is A and C pairwise independent? Answer Yes/No with justifications.

5. Suppose we are trying to maintain a random team of 11 players from a stream. Presently, 20 players have already passed, and the memory contains a uniformly random team of 11 players from the 20 that have passed. With what probability should we include the 21st player in the team? If we include the 21st player, how should we select the player to be removed?

6. There is a stream that contains m numbers from $\{1, 2, 3, \ldots, n\}$. What is the space required so that we will be able to output the arithmetic mean of all the elements that have passed?

7. **(Coupon Collector Problem)**. Suppose you are trying win movie tickets by buying biscuit packets. Along with every biscuit pack, you get a coupon with a number from $\{1, 2, \ldots, n\}$. Once you get coupons with all n numbers, you can redeem it for a movie ticket. Assume that the probability of finding any given number in a biscuit packet is $1/n$. What is the expected number of packets you need to buy in order to get each of the n numbered coupons?

8. Suppose that we roll a standard fair die 100 times. Let X be the sum of the numbers that appear over the 100 rolls. Use Chebyshev's inequality to bound $\Pr(|X - 350| \geq 50)$.

9. Consider a hash function $h : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, k\}$. Suppose m randomly chosen values from $\{1, 2, \ldots, n\}$ are hashed. What is the probability that none of these values i hash to $h(i) = 1$?

10. There is a stream of data, where each element is from $\{1, 2, 3, \ldots, n\}$. Suppose you come up with a random variable Y that you can measure from the stream, that has the following properties.

    - You can maintain Y using $O(\log n)$ space.
    - $E(Y) = F_2$, where F_2 is the second frequency moment[1].
    - $\mathrm{Var}(Y) = 2F_2^2$.

    Given parameters ε, δ, we need to get an estimate $\tilde{F}_2$ for F_2 such that $P[|\tilde{F}_2 - F_2| \leq \varepsilon F_2] \geq 1 - \delta$. Explain how you can do it and how much space is required.

[1]Let f_i be the number of times that element i appears in the stream. Then $F_2 = \sum_{i=1}^{n} f_i^2$ is the second frequency moment. For solving the problem, the definition of second frequency moment is not essential.