HW03

MATH 152, WINTER 2019

DUE WEEK 5-6

Reference: Foundations of Data Science by Blum, Hopcroft, and Kannan [BHK]

Reading: Chapter 6 [BHK] and lecture notes. Review the uniform distribute, and sampling techniques from your statistics courses.

1. (The uniform distribution). Let X be a uniform random variable over $[a, b]$. Find $\mathbb{E}[X]$ and $\text{Var}[X]$.

2. Let $X_1, X_2, \ldots, X_k$ be independent uniform random variables over $[0, 1]$, and let $Y = \min(X_1, X_2, \ldots, X_k)$. Show that

$$\mathbb{E}[Y^2] \leq \frac{2}{(k+1)^2}.$$

   (Hint: see the lecture note for the probability density function of Y.)

3. (Mean of estimator technique) Let X be any random variable with mean 0 and variance 1.

   (a) For $\epsilon \in (0, 1)$, use Chebyshev's inequality to bound $\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon)$.

   (b) The bound you get from Part a) is not useful for small ϵ (why?). Let's consider n independent copies $X_1, X_2, \ldots, X_n$ of X. That is, $X_1, X_2, \ldots, X_n$ are independent random variables which have the same distribution as X. Let

$$Z = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

   Use Chebyshev's inequality to bound $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \epsilon)$.

4. Let f_i be the frequency of the i symbol in the data stream. Let X_i be independent random variables assuming values 1 with probability $1/2$ and -1 with probability $1/2$. Let $V = (\sum_{i=1}^{m} X_i f_i)^2$. Show that $\mathbb{E}[V^2] \leq 3\mathbb{E}[V]^2$.

5. (Random Load Balancing[1]) Suppose you are a content delivery network – say, YouTube. Suppose that in a typical five minute time period, you get a million content requests, and each needs to be served from one of your, say, 1000 servers. How should you distribute the requests (let's call them "jobs") across your servers to balance the load? You might consider a round-robin policy, or a policy wherein you send each job to the server with the lowest load. But each of these requires maintaining some state and/or statistics, which might cause slight delays. You might instead consider the following extremely simple and lightweight policy, which is surprisingly effective: assign each job to a random server.

   Suppose n is the number of jobs and k the number of servers (e.g., $n = 106$ and $k = 103$). Define an indicator random variable X_j^i which will be 1 if the job j is assigned to server i and 0 otherwise. Then $X_j = \sum_{i=1}^{n} X_j^i$ denotes the load on machine i.

   (a) Find $\mathbb{E}[X_j]$.

[1] https://www.cs.princeton.edu/courses/archive/fall09/cos521/Handouts/probabilityandcomputing.pdf

(b) For each j, use Chernoff's inequality to bound

$$\mathbb{P}\left(X_j > \frac{n}{k} + 3\sqrt{\ln k}\sqrt{\frac{n}{k}}\right).$$

(c) Let $M = \max(X_1, \ldots, X_k)$. Find a lower bound for the following probability

$$\mathbb{P}\left(M \leq \frac{n}{k} + 3\sqrt{\ln k}\sqrt{\frac{n}{k}}\right).$$

6. (Approximating the mean). Given $\epsilon > 0, \delta \in (0,1)$, and a random variable X assuming values in $[0,1]$. Note that we don't know the distribution of X. But we want to estimate its mean, $\mathbb{E}[X]$, up to an error ϵ and with probability at least $1-\delta$. How can we do that? (Hint: Take $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ where X_i's are independent copies of X.)